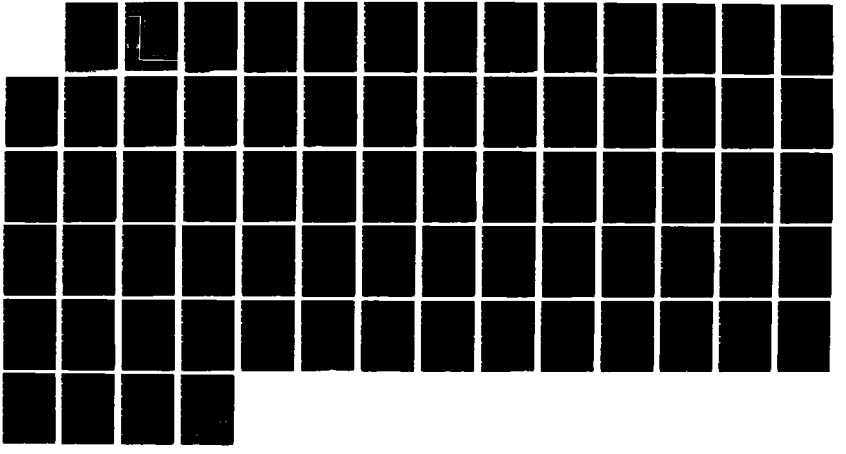
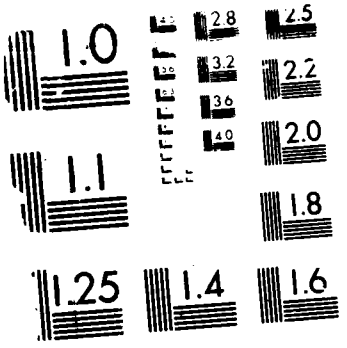
AD-A188 518 EXPERT SYSTEMS THEIR IMPACT ON PERFORMANCE AND COGNITIVE STRATEGIES IN DI (U) SOUTHEASTERN CENTER FOR ELECTRICAL ENGINEERING EDUCATION INC S S E GORDON 1/1

UNCLASSIFIED NOV 87 AFHRL-TR-87-13 F49620-82-C-0035 F/G 12/5 NL





MICROCOPY RESOLUTION TEST CHART

AFHRL-TR-87-13

(2)

AIR FORCE HUMAN RESOURCES LABORATORY

EXPERT SYSTEMS: THEIR IMPACT ON PERFORMANCE AND COGNITIVE STRATEGIES IN DIAGNOSTIC INFERENCE

Sallie E. Gordon

Department of Psychology
University of Idaho
Moscow, Idaho 83843

LOGISTICS AND HUMAN FACTORS DIVISION
Wright-Patterson Air Force Base, Ohio 45433-6503

November 1987

Final Report for Period January - December 1985

AIR FORCE SYSTEMS COMMAND
BROOKS AIR FORCE BASE, TEXAS 78235-5601

NOTICE

When Government drawings, specifications, or other data are used for any purpose other than in connection with a definitely Government-related procurement, the United States Government incurs no responsibility or any obligation whatsoever. The fact that the Government may have formulated or in any way supplied the said drawings, specifications, or other data, is not to be regarded by implication, or otherwise in any manner construed, as licensing the holder, or any other person or corporation; or as conveying any rights or permission to manufacture, use, or sell any patented invention that may in any way be related thereto.

The Public Affairs Office has reviewed this report, and it is releasable to the National Technical Information Service, where it will be available to the general public, including foreign nationals.

This report has been reviewed and is approved for publication.

ROSEMARIE J. PREIDIS
Contract Monitor

BERTRAM W. CREAM, Technical Director
Logistics and Human Factors Division

HAROLD G. JENSEN, Colonel, USAF
Commander

Unclassified

SECURITY CLASSIFICATION OF THIS PAGE AD-A188 518

REPORT DOCUMENTATION PAGE

Form Approved OMB No. 0704-0188

1a. REPORT SECURITY CLASSIFICATION Unclassified	1b. RESTRICTIVE MARKINGS
2a. SECURITY CLASSIFICATION AUTHORITY	3. DISTRIBUTION/AVAILABILITY OF REPORT Approved for public release; distribution is unlimited.
2b. DECLASSIFICATION/DOWNGRADING SCHEDULE	
4. PERFORMING ORGANIZATION REPORT NUMBER(S)	5. MONITORING ORGANIZATION REPORT NUMBER(S) AFHRL-TR-87-13
6a. NAME OF PERFORMING ORGANIZATION Southeastern Center for Electrical Engineering Education 6b. OFFICE SYMBOL (*If applicable*)	7a. NAME OF MONITORING ORGANIZATION Logistics and Human Factors Division
6c. ADDRESS (*City, State, and ZIP Code*) 1101 Massachussetts Avenue St. Cloud, Florida 32769	7b. ADDRESS (*City, State, and ZIP Code*) Air Force Human Resources Laboratory Wright-Patterson Air Force Base, Ohio 45433-6503
8a. NAME OF FUNDING/SPONSORING ORGANIZATION Air Force Office of Scientific Research 8b. OFFICE SYMBOL (*If applicable*) AFOSR	9. PROCUREMENT INSTRUMENT IDENTIFICATION NUMBER F49620-82-C-0035
8c. ADDRESS (*City, State, and ZIP Code*) Bolling Air Force Base, Washington, DC 20332	10. SOURCE OF FUNDING NUMBERS

PROGRAM ELEMENT NO	PROJECT NO	TASK NO	WORK UNIT ACCESSION NO
62205F	3017	08	04

11. TITLE (*Include Security Classification*)
Expert Systems: Their Impact on Performance and Cognitive Strategies in Diagnostic Inference

12. PERSONAL AUTHOR(S)
Gordon, S.E.

13a. TYPE OF REPORT	13b. TIME COVERED	14. DATE OF REPORT (*Year, Month, Day*)	15. PAGE COUNT
Final	FROM Jan 85 TO Dec 85	November 1987	72

16. SUPPLEMENTARY NOTATION

17. COSATI CODES		
FIELD	GROUP	SUB-GROUP
05	08	
09	04	

18. SUBJECT TERMS (*Continue on reverse if necessary and identify by block number*)
automation impacts
cognitive strategies
diagnostic inference
expert systems

19. ABSTRACT (*Continue on reverse if necessary and identify by block number*)

This report covers a 1-year research effort designed to develop a laboratory "sequential diagnostic inference" task, and assess the impact of introducing a computerized expert-aid system on performance of that task. The report details the laboratory task that was developed to be performed on a microcomputer, the expert system that assisted in the inference task, and the results of an initial experiment where subject performance was tracked under both manual and expert-aided conditions. Assessment of subject performance included a number of both subjective and objective measures such as accuracy, time to perform the task, subjective certainty, cognitive strategies, etc. Results indicate that: (a) the development of the laboratory inference task was successful, and the various task manipulations had a great impact on performance measures; (b) subjects' performance substantially changed as a result of using the expert system; and (c) subjects' strategies were identifiable and did not qualitatively change as a function of introducing the expert system.

Implications of the findings for the design and utilization of expert systems are discussed, as well as directions for future research.

20. DISTRIBUTION/AVAILABILITY OF ABSTRACT ☒ UNCLASSIFIED/UNLIMITED ☐ SAME AS RPT ☐ DTIC USERS	21. ABSTRACT SECURITY CLASSIFICATION Unclassified
22a. NAME OF RESPONSIBLE INDIVIDUAL Nancy J. Allin, Chief, STINFO Office	22b. TELEPHONE (*Include Area Code*) (512) 536-3877 22c. OFFICE SYMBOL AFHRL/TSR

DD Form 1473, JUN 86 *Previous editions are obsolete.* SECURITY CLASSIFICATION OF THIS PAGE

Unclassified

SUMMARY

The effects that expert systems may have on human decision making and task performance are not well known. Expert systems are computer programs designed to solve well-defined problems (e.g., selecting a course of action among several clearly delineated alternatives).

This research attempted to identify some of the interactions between a group of individuals and an expert system. The subjects had to identify animals using a set of clues. The expert system was designed to recommend the identities of the animals using nearly the same information that the subjects had. The experiment tested the performance of the subjects with and without expert-aiding. The experiment also tested performance of experienced versus inexperienced subjects when using the expert.

The findings indicated that the experienced subjects relied on the expert system even when they performed better without it; however, they relied more on the expert for the difficult identifications than for the easier ones. The inexperienced subjects consistently relied on the expert for animal identifications. They rated their abilities and that of the expert equally while the experienced subjects were more discriminating in their use of the expert. The subjects also displayed a tendency to request more information about the animals when using the expert than when they did not use it. The problem-solving strategies used by the subjects did not change when an expert system was introduced.

Although these findings were tentative and additional research is needed, a major conclusion was the importance of training. System operators should become well acquainted with their responsibilities, job requirements and their own capabilities before being introduced to expert system aids. This will allow them to more effectively use and assess the value of the expert system.

PREFACE

The objective of this research was the assessment of how humans perform diagnostic inference, and how a computer-aiding device impacts that performance.

This study was supported by the Air Force Office of Scientific Research under contract #F49620-82-C-0035 with the Southeastern Center for Electrical Engineering Education. The project was the result of work performed by the author during a Summer Faculty Research Fellowship at the Air Force Human Resources Laboratory (AFHRL), Wright-Patterson Air Force Base, Ohio.

Appreciation is expressed to Ms Rosemarie J. Preidis for providing guidance to meet Air Force needs, yet providing the freedom to perform research in an area of great personal interest. The author would also like to thank Mitch Sonnen, a student at the University of Idaho, who did an excellent job of programming the inference task and expert system. Special thanks to Ms Kathleen Y. Moorer for administrative support.

TABLE OF CONTENTS

	Page
I. INTRODUCTION	1
II. EXPERIMENTAL DESIGN	4
Task Characteristics	4
Manual Task Scenarios	6
The Expert System	9
Experimental Design	9
Subjects	13
Procedure	13
III. EXPERIMENTAL FINDINGS	14
Phase I:	
Manual Diagnostic Inference	14
Performance Measures	14
Performance Estimates	17
Strategy Classification	17
Data Analysis	21
Phase II:	
Expert-Aided Diagnostic Inference	30
Effects of Expert-Aid on Performance	30
Strategy Analysis	38
IV. SUMMARY AND DISCUSSION	45
Performance	45
Manual Conditions	45
Manual vs. Expert-Aiding Conditions	46
Strategy Analysis	49
Strategy Analysis for Manual Condition	49
Strategy Analysis for Expert-Aiding Condition	50
V. CONCLUSIONS	51
REFERENCES	52
APPENDIX A: POST-EXPERIMENTAL QUESTIONNAIRES	55
APPENDIX B: ANOVA TABLES FOR PHASE I	57
APPENDIX C: ANOVA TABLES - PHASE I VS. PHASE II	59
APPENDIX D: ANOVA TABLES - EXPERIENCE VS. NOVICE	62

LIST OF FIGURES

FIGURE		PAGE
1	Implementation of Computer-Aiding System	3
2	Displays for Manual Diagnostic Inference	7
3	Displays for Expert-Aided Diagnostic Inference	10
4	Mean Percent Correct as a Function of Experience and Session	34
5	Mean Time to Perform the Task as a Function of Attributes Available and Diagnosticity	35
6	Mean % of Trials the Expert was Asked as a Function of Subject Experience and Diagnosticity	37

LIST OF TABLES

TABLE		PAGE
1	Cell Means for Phase I (All dependent variables)	16
2	Percentages of First Attribute Requests-- Categorized According to Type of Split	22
3	Frequency of Second Attribute Requests-- Categorized According to Type of Split	26
4	Percentage of Favorite Attribute Requests for First and Second Requests (Sessions 2 and 3)	28
5	Percentage of First and Second Requests Falling Into Unique, Equal, and Non-Unique Categories	29
6	Cell Means for Experienced and Novice Subjects	31
7	Cell Means for Use of Expert System	32
8	Percentage of First Attribute Requests-- Categorized According to Split (Phase II)	39
9	Frequency of Second Attribute Requests Categorized According to Split (Phase II)	43
10	Percentage of Favorite Attribute Requests for First and Second Requests (Phase II)	44

I. INTRODUCTION

Computerized automation is becoming increasingly prevalent in a wide variety of positions in the armed services. This is especially true in the world of command, control and communications (C3), where much of the work involves complex "diagnostic inference." Diagnostic inference refers to a task where a person has informational cues, and on the basis of those cues, must infer the nature of the underlying cause or phenomenon. As technological complexities increase, human operators will have a more difficult time trying to understand, integrate, and utilize the information made available to them. In contrast to man's limited cognitive capacities and well documented biases [1,2], a computer can utilize and aggregate large volumes of information using pre-determined optimal strategies. It is no longer a question of whether computer aiding will be used, but how it will be used.

Just as there are problems inherent in using a completely "manual" system to perform these functions, there are also problems in using a completely "automated" system. These problems have been discussed at length elsewhere [2,3], but let it suffice to say that at the current time, expert systems are not sufficiently advanced to make automated systems infallible or able to deal with the multitude of unforeseen occurrences that are likely in the C3 environment.

Since neither human nor machine is solely capable of performing situational assessment functions, the solution lies in using both together and relying on the strengths of each. To integrate a person and machine successfully for a given task, one must understand how the human perceives and performs the task and analyze the best way to combine the capabilities of man and machine.

In order to optimally integrate human and machine, we need more information concerning two vital questions: (a) What factors influence the optimal performance of the task by the human and by the automation device? and (b) What factors determine operator acceptance and use of the automated system? For example, if the automation is extremely different from or incompatible with people's way of perceiving and accomplishing a task, then they may be less likely to accept and use the automation.

A review of the literature reveals that there is a large variety of computerized aiding systems being developed. Several of these aiding systems are specifically designed to aid in diagnostic inference types of tasks. For example, the PROSPECTOR [4,5] system helps geologists locate mineral deposits. MYCIN [6,7] and CADUCEUS [8] are systems which aid in medical diagnosis, and DENDRAL AND META-DENDRAL [9] analyze chemical data to make inferences about the structure of unknown chemical compounds. Although these systems are often referred to as "decision aids," the tasks fit into our definition of inference. In this type of system, the computer has a large data base of known facts or expert knowledge which is utilized when a new situation arises for analysis. The characteristics of the new situation are compared with the data base and an inference is generated.

Much of the work being done in this area is conducted by computer scientists and "knowledge engineering" experts [10,11]. This work involves two problems in expert systems: the knowledge base and the inference mechanism. Development of the knowledge base is known as knowledge engineering, and the problem is how to best transfer the knowledge that experts have into the most usable form within the computer data base [11,12]. A second problem involves development of the best inference mechanism or "inference engine." A variety of very sophisticated algorithms are currently under development [13].

Researchers are also finally becoming aware of the need to study human-computer interface, with a focus on the operator who will be using the expert system [12,14,15,16]. A volume recently edited by Salvendy [12] contains numerous papers concerned with user interface and acceptance of the automated system. Unfortunately, much of this work is concerned with the literal human-computer interface; that is,the language used, query system and so forth. There has been very little systematic research on the question of how human and machine are interfacing at the deeper task level [for exceptions see 17,18,19]. Some researchers have considered the importance of physician acceptance of the new diagnostic aiding systems[14,15]. Shortliffe [16] provides a list of factors that may influence a physician's decision to use the system; he has also suggested that even a highly reliable system may face difficulty in user acceptance [20].

Finally, Fitter and Cruikshank [21] obtained videotape data for three physicians: 59 consultations WITHOUT a computer system and 93 consultations WITH the system. Although the researchers assessed many interesting facets of the human-computer system, they did not make any attempt to systematically describe or measure the inference process used by the doctors before and after implementation of the system. Their only comment in this regard was that "the doctors appear to be influenced only to a minor extent by the feedback of the disease probabilities; they make very little use of feedback during the consultation but tend to check it against their own judgement at the end" (page 252).

It was felt that empirically derived data could be obtained to address the question of how the implementation of an expert system affects the performance of the human operator. A preliminary model was developed of the characteristics of the interaction between the human and the expert system. This model is applicable only to a situation where the human is completely in control of the task and uses the expert system as an aid, and is therefore free to consult (or not consult) the system and disregard the answer given by the expert system.

The system model is visually presented in Figure 1, with input to the system presented on the left. It can be seen that some situational variables will affect only the human operator's inference process, some will affect both human and expert system,and some will affect the operator's decision to accept the machine answer. Certainly there are other variables which would influence the process in specific task domains; however, it is felt that the variables listed represent those which are probably characteristic of most diagnostic

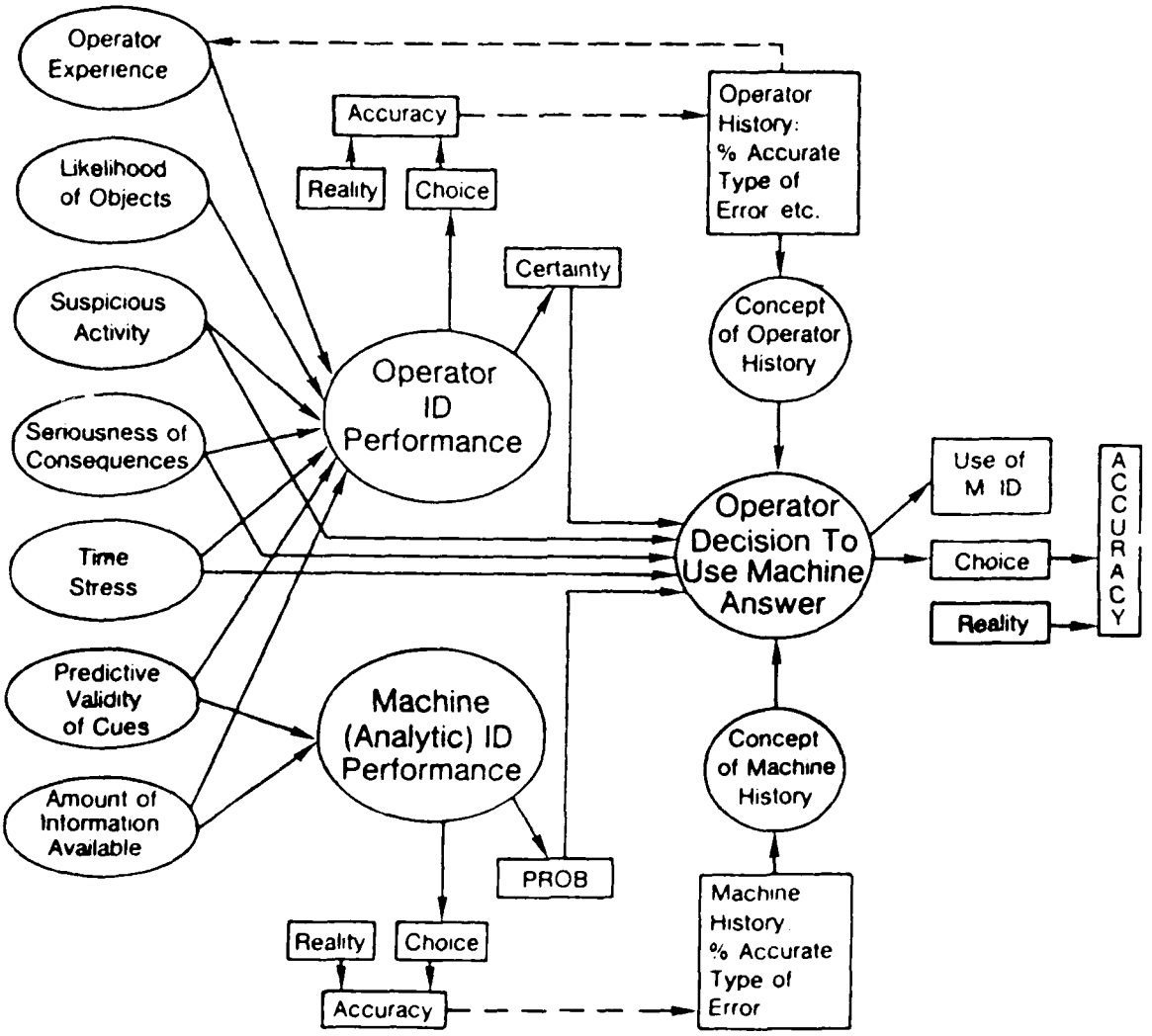


FIGURE 1. Implementation of Computer-Aiding System

inference tasks. Not represented here are factors that will be unique to an individual at the time of each inference. Primarily, these are "cognitive set" variables, where the specific information or attributes received will trigger a specific case in the history of the operator (maybe a recent or very common one). This is a process internal to the operator that is an interaction between the attribute set and the history of the particular operator. An example of this process would be a doctor seeing a patient who is having numbness on one side of the body. The doctor might hypothesize the cause as a stroke because (a) he just had a case similar to this yesterday, or (b) that is the most common cause of the symptom.

Aside from this analogical mechanism, we might expect that to some degree, the operator acts in a rational manner; that is, he considers attributes and searches his memory for causes which have matching attributes. To the extent that the subject is experienced and can rely on that memory, he will have confidence in his ability to draw the inference. It is felt that the operator will know when he is definitely certain of the answer or when the inference is tenuous. Especially in the latter case, the operator will consult the expert system for advice. Thus, factors causing the human to rely on the expert include (a) small amounts of incomplete information available, (b) low informativeness or predictive validity of the attributes, (c) time stress, (d) perceptions of the person's ability, (e) perceptions of the expert system's ability, and (f) seriousness of the consequences - the more serious the consequence, the less likely the person is to blindly accept the answer of the expert system.

This model guided the experimental design of the present research effort. A laboratory diagnostic inference task was developed such that several of the input variables could be manipulated while subjects performed the task unaided, and also while they were given the option of consulting an expert system.

II. EXPERIMENTAL DESIGN

Task Characteristics

The task developed was that of inferring an animal on the basis of a set of characteristics about the animal. An interactive program was written in Turbo Pascal for subjects to perform the task on an IBM personal computer. At the start of a trial, subjects were given an attribute describing an animal. They were then allowed to ask for information regarding other attributes. When subjects felt comfortable with giving a guess, they did so, and the trial ended. Under some conditions, the subjects performed the task by themselves; under other conditions, subjects were given the opportunity to use an "expert" built into the computer on which they were working. Before describing the task, it should be noted that much thought was given to the decision of using "existing" knowledge sets involving the real world, versus developing a new and artificial set of knowledge that the subjects learn before performing the task. After preliminary development of both kinds of tasks, it was decided that giving subjects a completely random and arbitrary knowledge base, while

being free from previous subject biases, would also be unrealistic and could easily cause cognitive processing different from that found in most real-life tasks.

In order to dampen the effects of their knowledge of animals and/or subject biases, subjects were "taught" the characteristics of eight animals at the beginning of the first session. Each animal was described in terms of six attributes:

1) Size (Large or Small)
2) Location (whether found in a Tree or on the Ground)
3) Speed (Fast or Slow)
4) Color (Brown or Grey)
5) Noise (whether the animal makes noise when traveling; Noise or No Noise)
6) Alarm (whether the animal sounds an alarm for approaching predators; Alarm or No Alarm).

The attributes taught to subjects for each of the eight animals are listed below. (We realize that the hawk and owl are not animals, but we will refer to them as such for the sake of brevity.)

RABBIT	GROUNDHOG	SQUIRREL
Small	Small	Small
Ground	Ground	Tree
Fast	Slow	Fast
Grey	Brown	Brown
No Noise	No Noise	Noise
No Alarm	No Alarm	No Alarm

DEER	OWL	HAWK
Large	Small	Small
Ground	Tree	Tree
Fast	Fast	Fast
Brown	Grey	Brown
No Noise	No Noise	No Noise
No Alarm	Alarm	Alarm

BEAR	WOLF
Large	Large
Ground	Ground
Slow	Fast
Brown	Grey
Noise	No Noise
Alarm	Alarm

In the following sections, the inference task will be described in detail. The research was conducted in two phases: one where subjects

performed the task under manual conditions (no expert system), and a second phase where subjects performed the same task and had the option of consulting an expert.

Manual Task Scenario

In the "manual" condition, subjects performed the inference task without the aid of an expert system. This phase was primarily designed to assess the cognitive processes and strategies being used by subjects. At the start of the trial, subjects were given information regarding one attribute (i.e., the animal is "small"). Then subjects were asked to provide a preliminary guess of the type of animal. This was simply a way of measuring the subjects' hypotheses at this point.

Figure 2a gives an example of the display screen as it was first presented to subjects. It can be seen that the eight possible answers were always listed at the top of the screen. The "trial" was simply a consecutive running number that informed subjects of the trial number for that session. The "condition" variable always read either "DAY" or "NIGHT." Subjects were told that if it were daytime, they would receive more information than if the trial occurred during the night. Finally, it can be seen that the attribute first given to this subject was "Small."

After entering their first guess, subjects were asked to give a certainty rating on a scale of 1 to 9, with 1 = not at all certain and 9 = extremely certain. Next, subjects were given a choice as to whether they wished to acquire more information or go on to the next trial. If they chose to continue, subjects were then allowed to ask for information regarding any of the remaining five attributes. As they asked for the information, their choices and the time to perform the choice were recorded. They were always required to give a guess and certainty rating after each new piece of information. The format of the question/answer interface is shown in Figure 2b. In the example given, the subject has just decided to ask about the animal location. Figure 2c shows the screen after the subject has acquired enough information to make a final guess. When the subjects have made their final guesses, they press "B" rather than "A" to signal that they are ready to go on to the next trial. At that point, they are informed as to the correct answer for that trial (see Figure 2d).

In the "Expert-Aid" condition, subjects performed basically the same as in the manual condition; however, they were allowed to ask an "expert" for help. This expert was built into the computer system (an integral part of the task program), and required only slight modification to the task. Subjects were allowed to consult the expert at any time during the trial, and were allowed to ask only twice per trial. The task scenario will be described in this section, and the mechanics of the expert system will be described in the following section.

It was decided that asking subjects to overtly hypothesize an animal after every attribute was not necessary in the Expert-Aid phase of the study. After

```
        POSSIBLE CHOICES

      RABBIT                              DEER
      SQUIRREL                            WOLF
      GROUNDHOG                           BEAR
      HAWK                                OWL

   TRIAL      :   1
   CONDITION  :   NIGHT


   CODE--ATTRIBUTES  :

 1--SIZE          :   Small
 2--LOCATION      :
 3--SPEED         :
 4--COLOR         :
 5--NOISE         :
 6--ALARM         :

   GUESS: type the FIRST  LETTER  of the animal.
```

2a. Initial Display Screen

```
      GROUNDHOG                           BEAR
      HAWK                                OWL

   TRIAL      :   1
   CONDITION  :   NIGHT


   CODE--ATTRIBUTES  :

 1--SIZE          :   Small
 2--LOCATION      :
 3--SPEED         :
 4--COLOR         :
 5--NOISE         :
 6--ALARM         :

   GUESS: type the FIRST  LETTER  of the animal.
R
   CONFIDENCE  RATING:   1 to 9,
2
   CONTINUE :  enter " A " NEXT TRIAL: enter " B "
A
   REQUEST ANOTHER ATTRIBUTE:
   ENTER the NUMBER to the LEFT of the ATTRIBUTE.
2
```

2b. Display After First Attribute Request

FIGURE 2. Displays for Manual Diagnostic Inference

```
      POSSIBLE CHOICES

      RABBIT                              DEER
      SQUIRREL                            WOLF
      GROUNDHOG                           BEAR
      HAWK                                OWL

  TRIAL     :   1
  CONDITION :   NIGHT


  CODE--ATTRIBUTES  :

1--SIZE         :   Small
2--LOCATION     :   Unknown
3--SPEED        :
4--COLOR        :   Grey
5--NOISE        :
6--ALARM        :   Alarm

  GUESS: type the FIRST  LETTER  of the animal.
```

2c. Display After Three Attributes Requested

```
  Your final answer was  Owl

  The correct answer is  Owl

  PRESS ANY KEY TO CONTINUE TRIALS
```

2d. Final Display Screen

FIGURE 2. Displays for Manual Diagnostic Inference(cont'd)

acquiring each attribute, the subject was given three choices: ask for more information, make a guess or ask the expert. An initial screen state for this task is given in Figure 3a. It can be seen that, in general, the display characteristics are similar to those used under manual conditions. Figure 3b shows how a display looks after the subject has asked for several attributes and has also asked the expert (expert answer is displayed to the right of the "Expert"). In the example, the subject has just entered a guess.

The Expert System

Because the information base for the task was small and well defined, it was a simple task to build an "expert" for this particular domain. First, a data bank was written to include all of the animal names and associated characteristics. Thus, the computer expert had perfect knowledge of the attributes and associated animals. Next, a subroutine was written which "read" the screen displayed at the time the subject asked for help. The attributes presented were matched to the attribute sets in the expert memory system and when a match was found, that animal was presented as the answer. Because the attribute sets were often incomplete, the computer could come up with more than one animal that matched the particular set of attributes being displayed. The program was written such that the order of matching took place randomly, and therefore the "choice" between more than one possible answer was a random one. Notice that only one answer was provided to the subject, not all possible answers.

Experimental Design

PHASE I

In the manual phase of the research, the following independent variables were manipulated:

(1) Session (1,2,or 3).
(2) Likelihood of the animal (Common vs. Rare).
(3) Number of Attributes Available (two vs. four).
(4) Diagnosticity of Attribute set (low vs. high).
(5) Monetary Payoff (low vs. high).

The first variable, Session, consisted of having subjects perform the task on three consecutive days. The primary purpose was to ascertain the effects of practice and operator experience on performance and cognitive strategy.

The Likelihood of the animal was manipulated by telling subjects before performing the task that five of the animals (squirrel, owl, hawk, deer, and rabbit) would be relatively common, and more likely to be the answer than the other three (bear, wolf and groundhog), which would be relatively rare. In setting up the trials, the common animals were, on the average, three times as likely to occur as the rare animals.

```
        POSSIBLE CHOICES

        RABBIT                              DEER
        SQUIRREL                            WOLF
        GROUNDHOG                           BEAR
        HAWK                                OWL

     TRIAL      :   1
     CONDITION  :   NIGHT
     EXPERT     :


     CODE--ATTRIBUTES  :

    1--SIZE         :   Small
    2--LOCATION     :
    3--SPEED        :
    4--COLOR        :
    5--NOISE        :
    6--ALARM        :

    (A)--ATTRIBUTE    (E)--EXPERT     (G)--GUESS
```

3a. Initial Display Screen

```
        RABBIT                              DEER
        SQUIRREL                            WOLF
        GROUNDHOG                           BEAR
        HAWK                                OWL

     TRIAL      :   1
     CONDITION  :   NIGHT
     EXPERT     :          Owl


     CODE--ATTRIBUTES  :

    1--SIZE         :   Small
    2--LOCATION     :   Unknown
    3--SPEED        :
    4--COLOR        :   Grey
    5--NOISE        :
    6--ALARM        :

    (A)--ATTRIBUTE    (E)--EXPERT     (G)--GUESS
G
 GUESS : type the  FIRST  LETTER  of the animal.
C
  CONFIDENCE  RATING :  1 to 9,
8
```

3b. Display After Guess and Confidence Rating

Figure 3. Displays for Expert-Aided Diagnostic Inference

The amount of information for subjects to use on any particular trial was manipulated by varying the number of Attributes Available. This was accomplished, by specifying for each trial, the values of either two or four attributes that would be available to the subject if the subject requested them. The other remaining attributes would yield a display of "unknown" when the subject asked for that information. For example, a two-attribute trial might have only information concerning size and location available to the subject, and all other dimensions would be unknown. To keep this from becoming obvious to the subjects, a small number of filler trials were given, with either three or five attributes available.

A second method of varying the difficulty of the trial, orthogonal to the amount of information available, was to vary the Diagnosticity of the entire set of cues. For any given set of attributes that could be potentially available (either two or four), that set of information could be highly diagnostic--where there was only one possible answer, or it could be low in diagnosticity--where there was more than one possible answer. For each trial, attribute sets were developed such that if the subject obtained all available information, there was either only ONE possible answer (high diagnosticity) or TWO possible answers (low diagnosticity).

Finally, in an effort to manipulate the "seriousness of the consequence" input variable (see Figure 1), the Monetary Payoff or amount of money to be earned by the subject for good performance was varied. One group of subjects was promised $.50 per day for getting a least 70% of the trials correct. The other group of subjects was promised $3.00 per day for getting at least 70% of the trials correct. All subjects were also told that they would receive a bonus for simply coming to all three sessions. All variables except the last, monetary payoff, were within-subject variables. The general design of each of the three sessions was as follows:

	Low Diagnosticity	High Diagnosticity
Two Attributes	3 Common animals 1 Rare animal	3 Common animals 1 Rare animal
Four Attributes	3 Common animals 1 Rare animal	3 Common animals 1 Rare animal

For each of the three sessions, 16 trials were critical in the assessment of the effects of the independent variables. These 16 trials corresponded to those listed in the table above.

The trials were developed such that animals were counter-balanced across the four conditions listed above (i.e., deer was equally represented in all four cells). Two separate sets of trials were developed for replication purposes. The two sets were equivalent in all ways, except for the exact animal/attribute combinations.

In the first session, subjects learned the experimental task and then performed 16 critical trials plus three additional "filler" trials. In the

second and third sessions, subjects performed 24 trials (16 critical trials plus eight "filler" trials). To maintain consistency across trials and sessions, only the 16 critical trials were used in the data analysis.

Information was recorded for attributes requested, guesses, and certainty ratings. All of the behaviors were recorded as they occurred, preserving the order; and in addition, the time in seconds (down to the hundredth) was recorded for each behavior. Finally, subjects were asked to fill out a questionnaire at the end of the third session (see Appendix A1). Part of the questionnaire involved asking subjects to estimate their performance for each of the three sessions. This allowed the measurement of the following dependent variables:

(1) Accuracy (percent correct).
(2) Time to perform the task.
(3) Subjective certainty on each trial.
(4) Attributes requested.

PHASE II

In the second phase of the project, the task differed in two respects: subjects gave only one guess and certainty rating at the end of the trial; and subjects were allowed to consult the expert. Most of the independent variables were retained in Phase II; Likelihood of the animal, number of Attributes Available, and Diagnosticity of the attribute set.

Twelve of the subjects from Phase I were asked to return for two more sessions; six were from the high pay condition, and six were from the low pay condition. They were deliberately chosen to represent a variety of ability in terms of their performance in Phase I. These subjects will be referred to as the "Experienced" subjects. In addition, 12 "Novice" subjects were run in Phase II, thus creating the between-subjects variable of Experience.

The within-subjects variables were combined in the same way as in Phase I, resulting in 16 critical trials during each of the two sessions. All new trials were created such that the experienced subjects would not see any that they had previously performed. Because of the small amount of time required to perform the trials (with only one guess at the end), subjects performed 30 trials total for each session. To ensure equivalency across manual and expert-aided sessions, most of the data analysis involved only 16 critical trials. As in Phase I, animals were counter-balanced across the experimental conditions, and two sets of replications were used.

After finishing the second session, subjects were asked to fill out a questionnaire (see Appendix A2). The first two questions asked the subjects to rate their own performance on a scale of 1 to 20 (with 1 being "extremely inaccurate" and 20 being "perfect"), and to rate the expert's performance on the same scale of 1 to 20. Additional items asked subjects to estimate their accuracy on each of the 2 days, describe their strategy, and comment on the expert system.

The design allowed for the following dependent variables to be assessed:

(1) Accuracy (percent correct).
(2) Time to perform the task.
(3) Subjective certainty on each trial.
(4) Attributes requested.
(5) Number of times the expert was asked for an answer.
(6) Number of times (under what conditions) the expert answer was used.
(7) Subjective perception of the subject's performance.
(8) Subjective perception of the expert's performance.

Subjects

Subjects were 36 upper-level students enrolled at the University of Idaho, with mean age of approximately 21 years. The subjects were obtained by having several instructors announce the experiment in their classes. Most of the students came from either psychology or engineering classes. Because of the possible difference in the two subject populations, every attempt was made to spread them evenly across the between-subjects variable (high or low payoff).

Four of the subjects who started in Phase I did not complete their three sessions and were replaced with other subjects. Three subjects did not complete Phase II and were replaced as well. Finally, one subject in Phase II never asked for the expert's advice during any of the 60 trials. It was decided that this was not similar to what would be expected in a real-world situation, and the subject was replaced with a new one.

Subjects were instructed as to the nature of the experiment and treated in accordance with American Psychological Association ethical guidelines. Informed consent was obtained, and names were stricken from all records as soon as data collection was complete for the subjects.

Procedure

Subjects were run individually in a small room equipped with several tables, chairs and one IBM personal computer. The experimenter was present the entire time the subject performed the task; however, the experimenter was usually reading or working at a table behind the subject.

When the subjects arrived the first day, they were told that we were studying how people solve problems and that they were going to be playing a simple guessing game. The nature of the task was briefly outlined to the subjects, and they were shown a list of the eight animals (with the rare animals marked with an asterisk). The subjects were then given a list of the eight animals and associated attributes (similar to the list given previously in this report). Subjects were allowed to study the list as long as they wished, up to 10 minutes. Most subjects studied the list between 6 and 8 minutes. Subjects were then placed in front of the computer and a practice trial was called up on the screen. Subjects in the "Expert-Aid" condition

were introduced to the use of the expert at this time. The display was explained to the subjects, as well as the key to press for each desired action. Subjects were allowed to perform the practice trial and ask the experimenter questions about the task. They were told that they should perform the trials accurately, but also as quickly as possible since they were being timed. Subjects in the low-pay condition were told that they would receive $.50 per session for getting at least 70% right; subjects in the high pay condition were told that they would receive $3.00 per session for getting at least 70% right.

All subjects completed the trial sets within 1 hour. In Phase I, most subjects took at least 30 minutes to perform all the trials in a session. In Phase II, experienced subjects usually completed the trials within 30 minutes, whereas the novice subjects generally took 30 to 40 minutes.

After subjects were finished with the last session in the phase, they were asked to fill out the questionnaire described earlier. The experimenter was available during that time to answer questions the subjects had on any of the items.

In Phase I, all subjects were paid $20, regardless of performance on the trials. In Phase II, all subjects were paid $14 each, regardless of performance on the trials. After the final session and questionnaire, subjects were debriefed and asked not to discuss the experiment with other potential subjects.

III. EXPERIMENTAL FINDINGS

Phase I: Manual Diagnostic Inference

First phase results of the manual performance will be reported first. The results will be presented in three sections: The first deals with overall subject performance on the task, the second summarizes results for subjects' performance estimates, and the third will detail the analysis of cognitive strategies.

Performance Measures

A Multivariate Analysis of Variance (MANOVA) was performed on the data from 24 Phase I subjects. Because the between-subjects variable of "pay condition" did not significantly affect subjects' performances, it was decided that analysis efforts would concentrate only on those 12 subjects who eventually went on to participate in Phase II (six from high pay and six from low pay).

The independent variables (all within-subjects) for the MANOVA were Session (1 vs. 2 vs. 3), Attributes Available (two vs. four), and Diagnosticity of the cue set (low vs. high). Analysis did not include Likelihood of the Animal (common vs. rare) because there were too few trials in each of those categories; thus, all data were collapsed over that

variable. The dependent variables included Percent Correct, Time, Certainty, and Number of Attributes Requested. The statistical summaries for this analysis are presented in Appendix B. The findings will be described here along with cell means. Only those effects which were significant in the multivariate test as well as the univariate test will be reported.

Means for the first dependent variable, Percent Correct (out of four), are presented in Table 1. A main effect was found for Session, where subject accuracy improved from session 1 ($\bar{x}$=.68) to session 2 ($\bar{x}$=.78) to session 3 ($\bar{x}$=.83). In addition, a main effect was found for Diagnosticity, where trials having a high diagnosticity (ONE animal possible) resulted in better performance ($\bar{x}$=.87) than did trials having a low diagnosticity ($\bar{x}$=.65).

The second dependent variable, time to perform the trial (from time of initial presentation to time of final guess) was also affected by Session and Diagnosticity of the cue set. The time to perform the task fell from a mean of 85 seconds for the first session, to 65 seconds for the second session, to 58 seconds for the third session. The main effect of Diagnosticity resulted in longer trial times for the low diagnosticity conditions ($\bar{x}$=77) than for the high diagnosticity conditions ($\bar{x}$=62). Finally, an interaction between Session and Diagnosticity revealed that, over sessions, subjects improved their trial times much more significantly for the easy trials than for the more difficult trials.

Subjects were asked to rate how certain they were of their guesses at the end of each trial (on a scale of 1 to 9). The certainty ratings for the subjects varied for the two Attributes Available conditions. When there were two possible attributes for them to acquire, subjects gave a mean certainty rating of 6.47. When there were four attributes possible, subjects gave a mean certainty rating of 6.99. Although this difference is statistically significant, it can be seen that the differences between the means is not a particularly great one. Since the number of attributes provided to subjects was manipulated orthogonally to the difficulty of the trials (Diagnosticity),the greater confidence in the trials with four attributes available indicates that subjects were "lulled" into believing that they were more accurate when they had more information.

A greater difference in certainty ratings was caused by the Diagnosticity of the trials. The easier high diagnosticity trials resulted in a mean rating of 7.58, whereas the more difficult low diagnosticity trials resulted in a mean rating of 5.77. This second rating seems fairly high considering that for all of these trials, subjects, by definition, HAD to be guessing between at least two animals. Finally, an interaction between Session and Diagnosticity revealed that subjects' confidence in their guesses increased over sessions for the easy trials, and dropped slightly over time for the more difficult trials.

Finally, subjects were assessed on the number of attributes they requested after the first attribute was presented. Means for this variable are thus number of attributes requested out of a total of five possible. A main

Table 1. Cell Means for Phase I (All Dependent Variables)

	Session 1	Session 2	Session 3
PERCENT CORRECT			
Low Diagnosticity			
Two Attributes	.50	.67	.79
Four Attributes	.71	.69	.56
High Diagnosticity			
Two Attributes	.79	.87	.98
Four Attributes	.71	.89	.98
TIME			
Low Diagnosticity			
Two Attributes	85.7	77.4	63.2
Four Attributes	90.3	74.7	72.6
High Diagnosticity			
Two Attributes	89.2	50.5	49.1
Four Attributes	76.5	57.2	49.6
CERTAINTY			
Low Diagnosticity			
Two Attributes	5.48	5.36	5.40
Four Attributes	6.47	6.08	5.86
High Diagnosticity			
Two Attributes	7.09	7.70	7.80
Four Attributes	7.44	8.20	7.87
ATTRIBUTES REQUESTED			
Low Diagnosticity			
Two Attributes	3.4	3.6	3.5
Four Attributes	3.1	3.3	3.7
High Diagnosticity			
Two Attributes	2.7	2.4	2.6
Four Attributes	2.6	2.6	2.5

effect of Diagnosticity was found, where high diagnosticity of trials resulted in a mean of 2.6 attributes requested whereas the low diagnosticity trials resulted in a mean of 3.4 attributes requested. In addition, a Session by Diagnosticity interaction showed a learning effect, where subjects learned to ask for more attributes for those trials where it was necessary (low diagnosticity trials).

Performance Estimates

Upon completion of Phase I, subjects were asked on a questionnaire to estimate the percent of trials they answered correctly for each of the three sessions. These estimates were then compared to the actual Percent Correct scores. An analysis of variance showed that subjects underestimated their performance for all three sessions, $F(1,11)=4.67, p=.05$. The means are given below for actual vs. estimated performance for all three sessions.

	ACTUAL	ESTIMATED
SESSION 1	.68	.60
SESSION 2	.78	.72
SESSION 3	.83	.79

Strategy Classification

To perform the strategy analysis, subjects' questionnaires were first reviewed to determine what subjects thought that they were doing. This resulted in several categories of strategies, and the experimenter also tried to determine a reasonable classification of strategies, partly on the basis of previous research in this area. At this point, the raw data for subjects in sessions 2 and 3 were reviewed to determine whether evidence could be found for any of the strategies defined. On the basis of this data review work, the strategy classifications were slightly revised. The resultant classification system contains five cognitive strategies that subjects may have used during the course of the task. These strategies are given below (an example will depict each strategy):

(1) Half-Split. This is a well-known classical strategy that is the most "rational" procedure possible for the task [22]. After the initial attribute acquisition, the subject brings into "working memory" all of the animals having that attribute. Then a new attribute is requested which comes closest to "splitting" the set of possible animals into equal groups. This procedure is followed until the set is narrowed down to one animal or the subject runs out of attribute information. (This strategy is not the only means to the desired end; however, it ensures getting there with the fewest attribute requests.)

EXAMPLE:

First Attribute Given: NO NOISE

Subject:

(1) Determine possible animals:

RABBIT
HAWK
OWL
DEER
GROUNDHOG
WOLF

(2) Determine splits:

SIZE	Small(4)	Large(2)
LOC	Grnd(4)	Tree(2)
SPEED	Fast(5)	Slow(1)
COLOR	Grey(3)	Brwn(3)
ALARM	No (3)	Yes (3)

(3) Choose attribute resulting in closest to "equal" split: ASK FOR COLOR OR ALARM

Second Attribute Given: Grey

Subject:

(1) Determine possible animals:

RABBIT
OWL
WOLF

(2) Determine splits:

SIZE	Small(2)	Lge(1)
LOC	Grnd (2)	Tree(1)
SPEED	Fast (3)	Slow (0)
ALARM	No (1)	Yes (2)

(3) Choose attribute resulting in closest to "equal" split: ASK FOR SIZE, LOC, or ALARM

etc

(2) Set Reduction. This is an easier way to reduct the alternatives than the first strategy. Again, after the first attribute, the subject must think of a set of animals with the attribute. However, the subject need think only of a set that can be differentiated on SOME other attribute. The subject

does not have to think of ALL animals with that attribute, nor then consider ALL possible attribute requests and whether one is best. The subject will simply determine an attribute that is diagnostic to any degree at all, and request that attribute. The subject then determines a set of animals that fit the two attributes known, and determines whether there is another attribute that can differentiate among them. This continues until the subject narrows down the set enough to ensure consideration of all possible animals with the known attributes (which may not happen at first). A reduction of the set to two alternatives will place the subject in the same position as for the Half-Split strategy.

EXAMPLE:

First Attribute Given: NO NOISE
Subject:

(1) Determine some set
of animals with known
attribute: RABBIT
WOLF
OWL

(2) Choose any attribute
which will differentiate
among these: ASK FOR LOC

Second Attribute Given: GROUND

Subject:

(1) Determine some set
of animals with known
attributes: RABBIT
DEER
GROUNDHOG

(2) Choose any attribute
which will differentiate
among these: ASK FOR SIZE

Third Attribute Given: LARGE

Subject:

(1) Determine some set
of animals with known
attributes: DEER
WOLF

(2) Choose an attribute which will differentiate among animals: ASK FOR ALARM

etc

(3) Hypothesis Testing. In this strategy also, the subject considers some subset of animals that have the initial attribute. However, the subject will clearly have a favorite (hypothesis) and will request an attribute to confirm that choice. That is, they will ask for an attribute that characterizes that animal and (optimally) no other.

EXAMPLE:

First Attribute Given: NO NOISE

Subject:

(1) Determine some of possible animals, with one favorite: GROUNDHOG (favorite)
RABBIT
WOLF

(2) Choose attribute to request that is most uniquely characteristic of favorite animal: ASK FOR SPEED (looking for slow)

Second Attribute Given: FAST
Subject:

(1) Determine new set with favorite: RABBIT (favorite)
HAWK
DEER

(2) Choose attribute to request that is most uniquely characteristic of favorite animal: ASK FOR COLOR (looking for grey)

etc

(4) Favorite Attributes. In this strategy, subjects did not attempt to reduce a set of possible alternatives, but rather, their attention was focused more on acquiring certain kinds of information. This strategy would lead to

having favorite attributes which are always requested first, second, third, and so forth. (Subjects might still keep track of possible animals to some degree; otherwise, they would always ask for the same number of attributes in the same order no matter what the trial, and this never seemed to occur.)

(5) Random Request. This is an unlikely but possible strategy where, after the initial attribute, the subject simply randomly chooses one of the remaining five attributes to request. (Notice that this is compatible with a general hypothesis testing strategy that makes no effort to assess the best attributes to request; an animal is hypothesized and any attribute might be requested to confirm that hypothesis.)

Data Analysis

To determine which of the five strategies subjects were using, it was necessary to derive predictions based on each strategy and compare the data with those predictions. The prediction and relevent data for each strategy will be described in this section. In obtaining data for the strategy analysis, only data from sessions 2 and 3 were included.

To begin, three of the strategies will be covered together because the same set of data is relevant to them.

(1) Half-Split:

The Half-Split strategy predicts that the attribute requested will be that which most evenly splits the possible animals into groups. Thus, for each beginning attribute it was possible to determine the "best" split(s). A second strategy of splits was determined which were acceptably diagnostic but not optimally so, and finally, a third category of attribute requests was identified which was not at all diagnostic. For example, if "Small" is the first attribute given, the best attribute requests are Location, Color, and Alarm. The acceptable attribute requests would be Speed or Noise. The Half-Split strategy predicts that subjects will always request the optimum attribute or the "best" split. Table 2 shows these three categories of attribute requests, and specifies the predicted attribute requests for each possible first attribute. The numbers which refer to subject data will be explained below.

(2) Set Reduction:

The Set Reduction strategy predicts that subjects will request any attribute that is at all diagnostic. In terms of the three types of attribute requests listed in Table 2, this strategy predicts that subjects will request the first or second type of attribute, but never the non-diagnostic attribute in the far right column. To summarize, the Half-Split predicts attribute requests only of the "best" split variety, whereas the Set Reduction strategy predicts use of both "best" and acceptable" split attribute requests.

Table 2. Percentages of First Attribute Requests--Categorized According to Type of Split

FIRST ATTR	"BEST" SPLIT REQUESTS	"ACCEPTABLE" SPLIT REQUESTS	"NON-DIAG" REQUESTS
Small	Loc .40	Speed .00	
	Color .18	Noise .16	
	Alarm .26		
	TOTAL .84	TOTAL .16	--
Large	Speed .09		Loc .07
	Color .27		
	Noise .18		
	Alarm .39		
	TOTAL .93	--	TOTAL .07
Ground	Size .31	Noise .21	
	Speed .17		
	Color .21		
	Alarm .10		
	TOTAL .79	TOTAL .21	--
Tree	Color .40		Speed .00
	Noise .23		Size .01
	Alarm .36		
	TOTAL .99	--	TOTAL .01
Fast	Loc .27	Size .23	
	Color .21	Noise .06	
	Alarm .23		
	TOTAL .71	TOTAL .29	--
Slow	Size .34		Color .07
	Noise .28		Loc .07
	Alarm .24		
	TOTAL .86	--	TOTAL .14
Brown	Size .36		
	Loc .19		
	Speed .14		

Table 2 (Concluded)

	"BEST" SPLIT REQUESTS	"ACCEPTABLE" SPLIT REQUESTS	"NON-DIAG" REQUESTS
	Noise .17		
	Alarm .14		
	TOTAL 1.00	--	--
Grey	Size .32		Noise .03
	Loc .37		Speed .00
	Alarm .28		
	TOTAL .97	--	TOTAL .03
No Noise	Color .16	Loc .32	
	Alarm .10	Size .38	
		Speed .04	
	TOTAL .26	TOTAL .74	--
No Alarm	Size .28		
	Loc .37		
	Speed .05		
	Color .18		
	Noise .12		
	TOTAL 1.00	--	--
Alarm	Size .34	Speed .04	
	Loc .24	Noise .17	
	Color .21		
	TOTAL .79	TOTAL .21	--
Mean Percentage OBTAINED for all Combined (N= 576):	(n = 479) .83	(n= 85) .15	(n= 12) .02
Mean Percentage EXPECTED based on Half-Split Strategy:	1.00	.00	.00
Mean Percentage EXPECTED based on Set Reduction Strategy:	.83	.17	.00
Mean Percentage EXPECTED based on Equal Choice:	.67	.16	.16

(3) Random Request:

This strategy predicts that the subjects will not discriminate among the three types of categories discussed above, and that the second attribute requests will spread equally across the alternative attributes. This means that the expected values for each of the categories can be obtained by determining the proportion of attributes occurring in each category.

The data for subjects' first attribute requests are given in Table 2. The numbers represent the percentages of requests that fell in a particular category. For example, when Small was the first attribute provided, subjects asked for Location in 40% of the trials. The total for each subset is given below the set of attributes in a given section (e.g., when Small was the first attribute, the "best" split attributes were requested on 84% of the trials). The percentages for all trials combined are given below the final TOTAL row.

It can be seen by comparing the obtained data with the predictions given at the bottom of Table 2 that the data are much more consistent with the Set Reduction strategy than the Half-Split strategy. However, they are not entirely consistent with the Set Reduction strategy because there were a few trials where subjects asked for a NON-DIAGNOSTIC attribute (category #3). Because the expected value for this category is zero, and the obtained value was greater than zero, a Chi-Square analysis for goodness-of-fit could not be conducted for either of the first two strategies. A goodness-of-fit test was conducted using the expected values based on an equal (proportional) use of the three categories (the last set of expected values in Table 2). A Chi Square analysis showed that subjects' requests did differ significantly from these three expected values, Chi-Square (2) =95.2, p=.001.

To try to determine whether the data fit either strategy (1) or (2) more closely, an analysis was conducted using the data for only those trials where subjects had a choice between "best" split attributes and "acceptable" split attributes. In this analysis, the Set Reduction strategy produced the expected values, based on the assumption that since subjects did not discriminate between "best" and "acceptable" attribute requests, the two would be chosen an equal amount of the time (or more accurately, proportionately to the number of attributes in each category). Thus, a test of the difference between the obtained values and the expected values was a test of the Set Reduction strategy. To the extent that the obtained values differed in the direction of the "best" split, this would provide indirect support for the Half Split strategy.

The obtained and expected values for each of the two categories were:

	"Best" Split	"Acceptable" Split
OBTAINED	140	85
EXPECTED (Set Reduction)	127	98

A Chi-Square test for goodness of fit showed that the obtained values did NOT significantly differ from the expected values based on the Set Reduction strategy, Chi-Square (1) =3.05. This can be taken as evidence in support of the Set Reduction strategy relative to the Half-Split strategy.

An analysis of second attribute requests was carried out in a manner similar to that just described. The data obtained for this analysis are more complicated because at this point, subjects have obtained information for two attributes (although in some cases, the second piece of information was "unknown"). Frequencies were calculated for all possible combinations of known attributes and the resultant second attribute requests (for example, one combination would be ALARM given first, NO NOISE given as the first request, then COLOR as the second request). These combinations were then categorized as to whether they were the "best split" possible, an "acceptable split," or a non-diagnostic request.

It turned out that for all combinations, there were no instances where a decision for the second request had to be made among all three types of categories. A decision had to be made among the following alternatives:

1) all second requests were category #1
2) all second requests were category #2
3) all second requests were category #1 or #2
4) all second requests were category #1 or #3.

An example of (4) would be SMALL and GREY given, with the alternative second requests being:

Location	Category #1 ("Best" Split)
Alarm	Category #1
Speed	Category #3 ("Non-Diagnostic")
Noise	Category #3.

The frequencies of choices made between categories 1 and 3 were used to compare the Half-Split and Set Reduction strategies with a simple Random Request strategy. The data for these frequency tabulations are given in the top half of Table 3. The frequencies are listed according to the choices possible. The example given above would fall under the "2 Best, 2 Non-diagnostic" section. Totals were obtained for all choices involving #1 vs #3 categories. These totals are listed at the bottom of the first section, along with the expected frequencies based on a random choice among the alternative second requests. A Chi-Square analysis was performed on these data and showed that the obtained frequencies were nowhere near what would be expected on the basis of random choice, Chi-Square (2)=182.2, $p<.001$. This provides support for both the Half-Split and Set Reduction strategies. (It was not possible to distinguish between the two in this analysis.)

A similar analysis was performed for the data comparing the frequency of choices between category #1 ("Best"Split) and category #2 ("Acceptable"

Table 3. Frequency of Second Attribute Requests Categorized According to Type of Split

CATEGORY 1 VS 3, CHOICE BETWEEN:	"BEST SPLIT" REQUESTS	"NON-DIAGNOSTIC" REQUESTS
3 Best, 1 Non-Diag.		
OBTAINED	35	3
EXPECTED	58.5	19.5
2 Best, 2 Non-Diag.		
OBTAINED	158	9
EXPECTED	83.5	83.5
1 Best, 3 Non-Diag.		
OBTAINED	28	2
EXPECTED	7.5	22.5
TOTAL OBTAINED	261	14
TOTAL EXPECTED (Random)	149.5	125.5

CATEGORY 1 VS 2 CHOICE BETWEEN:	"BEST" SPLIT REQUESTS	"ACCEPTABLE" SPLIT REQUESTS
3 Best, 1 Acceptable		
OBTAINED	28	4
EXPECTED	24	8
2 Best, 2 Acceptable		
OBTAINED	40	21
EXPECTED	30.5	30.5
1 Best, 3 Acceptable		
OBTAINED	9	18
EXPECTED	6.75	20.25
TOTAL OBTAINED	77	43
TOTAL EXPECTED (Set Reduction)	61.25	58.75

Split). These frequencies are given in the lower half of Table 3 and are also arranged according to the division of alternatives (three out of the four alternatives were the Best Split, etc.). Totals were obtained for all choices between category #1 and #2, and are listed at the bottom of the table. The Half-Split strategy predicts that category #1 would always be chosen over category #2, whereas the Set Reduction strategy predicts that no preference would be shown for either category type, and therefore the frequencies should be the same as those expected by chance. The expected frequencies based on the Set Reduction strategy were calculated and are given below the TOTAL OBTAINED values. First, it can be seen that the obtained values do not lend support to the Half-Split strategy predictions (which would predict only category #1 choices). However, a Chi-Square analysis showed that the obtained values did significantly differ from the expected values based on the Set Reduction strategy, Chi-Square(1)=8.27, $p < .01$. The difference is in the direction of a greater frequency of category #1 choices than would be expected.

4) Favorite Attributes:

To determine whether some (or all) subjects had favorite attributes that they requested regardless of the specifics of the trial, their first and second attribute requests were classified for sessions 2 and 3. These data are provided in Table 4. For a given subject, the most frequent first attribute requested was determined. The percentages given in the first two columns of the table represent how often each subject asked for the most requested attribute in sessions 2 and 3. The columns on the right represent the same type of data for the second attribute requested. For example, subject #11 had a favorite attribute that he chose for the first request 50% of the time in session 2 and 83% of the time in session 3. However, this subject did not have one consistent second request in both sessions 2 and 3.

The expected value for the first requests, given NO favorites would be 1/5 or 20%, and the expected value for second requests would be 1/4 or 25%. It can be seen that the majority of subjects seemed to have a favorite first and second attribute request, and half of the subjects used this strategy.

5) Hypothesis Testing Strategy:

As explained earlier, this strategy predicts that the subject will request an attribute that is the most unique characteristic of the animal hypothesized by the subject. It was possible to assess the existence of this strategy because subjects were asked to give a hypothesized animal after each attribute acquisition. Thus, for each initial attribute given, it was determined which attribute would be most "unique" for each possible hypothesis. The attributes which were NOT most unique were categorized as to whether they were equally descriptive of all possible alternatives, or "non-unique" (that is, they were characteristic of the favorite plus many other animals as well). A frequency count was determined for subjects according to whether their first attribute requested was "unique," "equal," or "non-unique." The percentages of attribute requests falling into each of these three types of categories are given in Table 5. A similar analysis was conducted for the second attribute requested. These are shown in the right half of the table.

Table 4. Percentage of Favorite Attribute Requests for First and Second Requests (Sessions 2 and 3)

	FIRST REQUEST		SECOND REQUEST	
Subject	Session 2	Session 3	Session 2	Session 3
1	Color .33	Color .29	Color .32	Alarm .38
2	Color .46	Size .33	Alarm .39	Alarm .33
3	Size .37	Size .29	Noise .36	Alarm .25
4	Noise .42	Alarm .37	Noise .27	Noise .35
5	Size .29	Size .29	Noise .26	Size .24
6	Loc .58	Loc .37	Alarm .43	Color .48
7	Size .46	Alarm .58	Loc .33	Noise .40
8	Alarm .37	Size .50	Color .32	Color .33
9	Alarm .29	Speed .25	Alarm .30	Alarm .25
10	Alarm .37	Noise .42	Alarm .40	Alarm .38
11	Color .50	Color .83	Noise .40	Loc .43
12	Size .46	Size .29	Color .28	Color .26
Mean OBTAINED FREQUENCIES	.41	.40	.34	.34
Mean EXPECTED FREQUENCIES (Random Choice)	.20	.20	.25	.25

Table 5. Percentage of First and Second Requests Falling Into Unique, Equal, and Non-Unique Categories

	FIRST ATTRIBUTE REQUEST			SECOND ATTRIBUTE REQUEST		
Subject	Unique	Equal	Non-Unique	Unique	Equal	Non-Unique
1	.36	.50	.69	.46	.26	.89
2	.27	.61	.46	.31	.56	.70
3	.46	.48	.31	.42	.42	.64
4	.33	.57	.46	.23	.45	.69
5	.39	.57	.36	.38	.50	.56
6	.29	.53	.64	.42	.54	.57
7	.42	.56	.33	.48	.30	.91
8	.50	.42	.31	.45	.37	.78
9	.51	.35	.61	.27	.64	.50
10	.50	.46	.40	.65	.25	.73
11	.42	.42	.75	.40	.48	.62
12	.65	.22	.64	.35	.48	.64
For Subjects Combined:						
Frequency	184	226	72	109	145	82
Total Poss.	431	475	145	275	332	119
Total %	.43	.48	.50	.40	.44	.69

If this strategy were descriptive of subjects' attribute requests, it would be expected that the attributes in the "unique" column would be chosen over the other types (equal or non-unique). It can be seen that this was clearly not the case; i.e., subjects did not tend to request the attribute that would be most unique to the hypothesized animal. Thus, we can be reasonably confident in ruling out this strategy as a likely description of subjects' cognitive processes.

The data reviewed thus far indicate support for BOTH a Set Reduction strategy AND a tendency to request certain attributes more than others. These two strategies are not at all incompatible, given that the Set Reduction strategy often allows flexibility as to which attribute should be requested (that is, many will often be diagnostic for a subset of alternatives).

Phase II: Expert-Aided Diagnostic Inference

Data Analysis from the second phase of the effort will be presented first for the various performance measures and then for the strategy analysis.

Effects of Expert-Aiding on Performance

Data for the first four dependent variables--Accuracy, Time, Certainty Ratings, and Attribute Requests are presented in Table 6. The data are organized according to phase, with Phase I included to assess the impact of introducing an expert system after the task has been learned. Each score for the dependent variables was based on three common and one rare animal. Data from two additional dependent variables--Percent Expert was Asked and Percent Expert was Used--are presented in Table 7; these data were collected from Phase II only.

The results were analyzed via two separate data analyses:

1. A MANOVA was first performed for the Percent Correct, Certainty, and Attribute Request data from the 12 subjects who performed in Phase I and Phase II. The independent variables were Phase (I vs. II), Session (2 vs. 3 for Phase I and 1 vs. 2 for Phase II), Attributes Available (2 vs. 4), and Diagnosticity (low vs. high). Results from the analysis are summarized in Appendix C.

2. A MANOVA was performed for the data from subjects in Phase II only (all dependent variables shown in Tables 6 and 7). The independent variables were Experience (Experienced vs. Novice), Session (1 vs. 2), Attributes Available (2 vs. 4), Diagnosticity (low vs high). Results from the analysis are presented in Appendix D.

For the sake of clarity, each dependent variable will be discussed with regard to all subjects, regardless of the specific analysis used. For the first variable, Percent Correct, a main effect was found where "experienced" subjects performed significantly better in Phase I, the manual condition ($\bar{x}$= .80) than they did with the expert-aid ($\bar{x}$=.74). In addition, experienced subjects performed significantly better than the novice subjects, even when

Table 6. Cell Means for Experienced and Novice Subjects

	Phase I Experienced		Phase II Experienced		Phase II Novice	
(Session)	2	3	1	2	1	2
PERCENT CORRECT						
Low Diag.						
2 Attributes	.67	.79	.58	.60	.48	.52
4 Attributes	.69	.56	.50	.48	.56	.37
High Diag.						
2 Attributes	.87	.98	.98	.96	.94	.94
4 Attributes	.89	.98	.85	.94	.69	.83
TIME						
Low Diag.						
2 Attributes			52	41	56	41
4 Attributes			50	45	56	41
High Diag.						
2 Attributes			38	31	49	34
4 Attributes			44	38	53	37
CERTAINTY						
Low Diag.						
2 Attributes	5.4	5.4	6.1	6.3	5.8	5.1
4 Attributes	6.1	5.9	6.5	6.1	5.8	5.7
High Diag.						
2 Attributes	7.7	7.8	8.4	8.4	7.0	7.7
4 Attributes	8.2	7.9	8.4	8.6	7.3	8.0
ATTRIBUTES REQUESTED						
Low Diag.						
2 Attributes	3.6	3.5	3.9	3.8	3.7	3.6
4 Attributes	3.3	3.7	3.7	3.9	3.6	3.7
High Diag.						
2 Attributes	2.4	2.6	2.9	2.8	3.0	2.9
4 Attributes	2.6	2.5	2.9	3.4	3.2	3.7

Table 7. Cell Means for Use of Expert System

	Phase II Experienced		Phase II Novice	
Session	1	2	1	2
PERCENT EXPERT ASKED				
Low Diag.				
2 Attributes	.87	.77	.60	.54
4 Attributes	.77	.57	.52	.48
High Diag.				
2 Attributes	.14	.04	.46	.21
4 Attributes	.29	.12	.33	.12
PERCENT EXPERT USED				
Low Diag.				
2 Attributes	.69	.56	.70	.54
4 Attributes	.61	.56	.65	.60
High Diag.				
2 Attributes	.33	.17	.57	.42
4 Attributes	.42	.42	.54	.25

comparing the first two manual sessions of the experienced subjects with the first two sessions of the novice subjects (means were .73 and .67, respectively), $F(1,22)=4.02$, $p=.05$. Figure 4 shows the overall pattern of data for both groups of subjects.

Evidence bearing on the reason for the drop in performance from manual to expert-aided conditions comes from an interaction found between Phase and Diagnosticity. That is, for the experienced subjects, performance stayed relatively high for the easy high diagnosticity trials (.93 for both Phase I and II). However, for the difficult low diagnosticity trials, performance actually decreased from .68 in Phase I to .54 in Phase II. Two alternative reasons for this effect will be given in the discussion section.

Two other effects were obtained for accuracy (Percent Correct) data. First, in both multivariate analyses, Diagnosticity of the cue set was found to have a large impact on the accuracy of subjects (overall, low diagnosticity items resulted in 57% correct whereas high diagnosticity resulted in 90% correct). A second finding was that for subjects in Phase II, the number of attributes available had an impact on accuracy (means were 75% and 65% for two and four attributes, respectively). This was opposite to what might be expected, and probably indicates that when four attributes were necessary to narrow the choice to one, subjects did not spend the effort to obtain all of the relevant information.

For the second dependent variable, Time to perform the trial, subjects did NOT differ based on whether they had previous manual experience in the task. Also, all subjects in Phase II were equally affected by the manipulation of task variables. A main effect was found for Session, where subjects decreased their time in general. In addition, there were main effects for Attributes Available (two attributes took about 42 seconds whereas four took 45 seconds) and Diagnosticity (easy high diagnosticity trials took 40 seconds whereas low diagnosticity trials took an average of 48 seconds). Finally, an interaction between Attributes Available and Diagnosticity is shown in Figure 5. It can be seen that the low diagnosticity trials took more time regardless of the number of attributes available, presumably because subjects had to request most or all of the information in either case.

The dependent variable of Certainty was affected by several factors. First, experienced subjects showed a high mean certainty of their answers in Phase II (7.3) than in Phase I (6.8). As expected, their ratings were also higher in Phase II than ratings given by novice subjects (see Table 6). In contrast to the Phase I results, subject certainty scores in Phase II were not significantly affected by the number of attributes available in the trial (see Appendix D). However, as previously found in Phase I, diagnosticity of the cue set strongly affected subject certainty. In addition, for all subjects in Phase II, a Session x Diagnosticity interaction occurred where subjects' certainty ratings for high diagnosticity trials increased from 7.8 to 8.2 over the two sessions, and ratings for the low diagnosticity trials decreased from 6.0 to 5.8 over the two sessions.

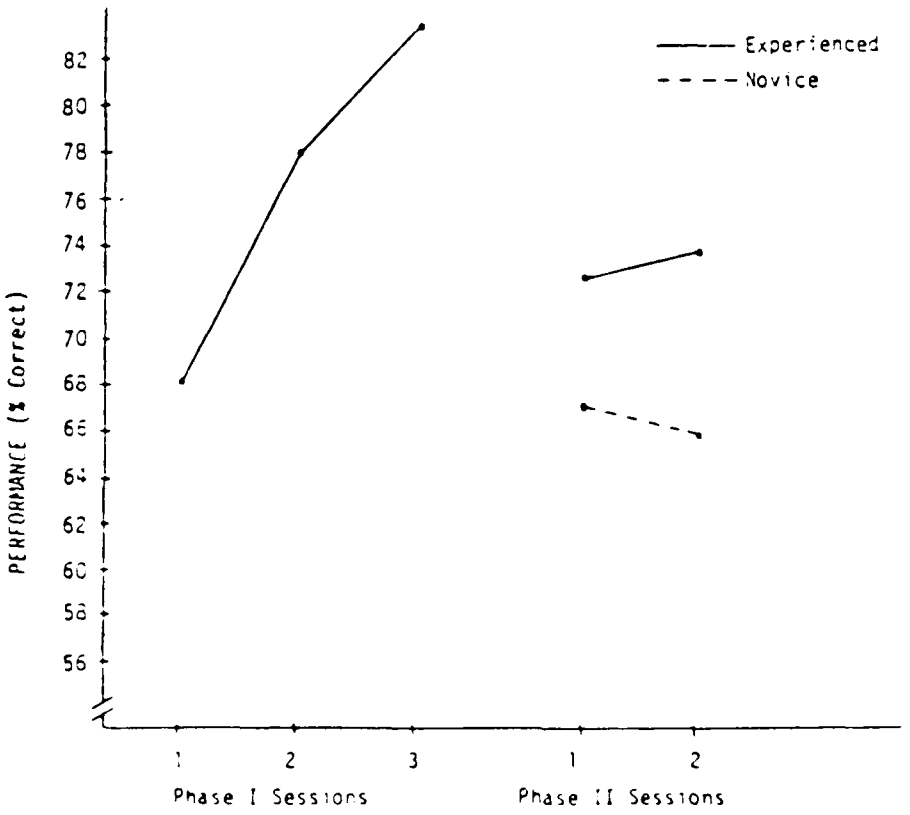


FIGURE 4. Mean Percent Correct as a Function of Experience and Session

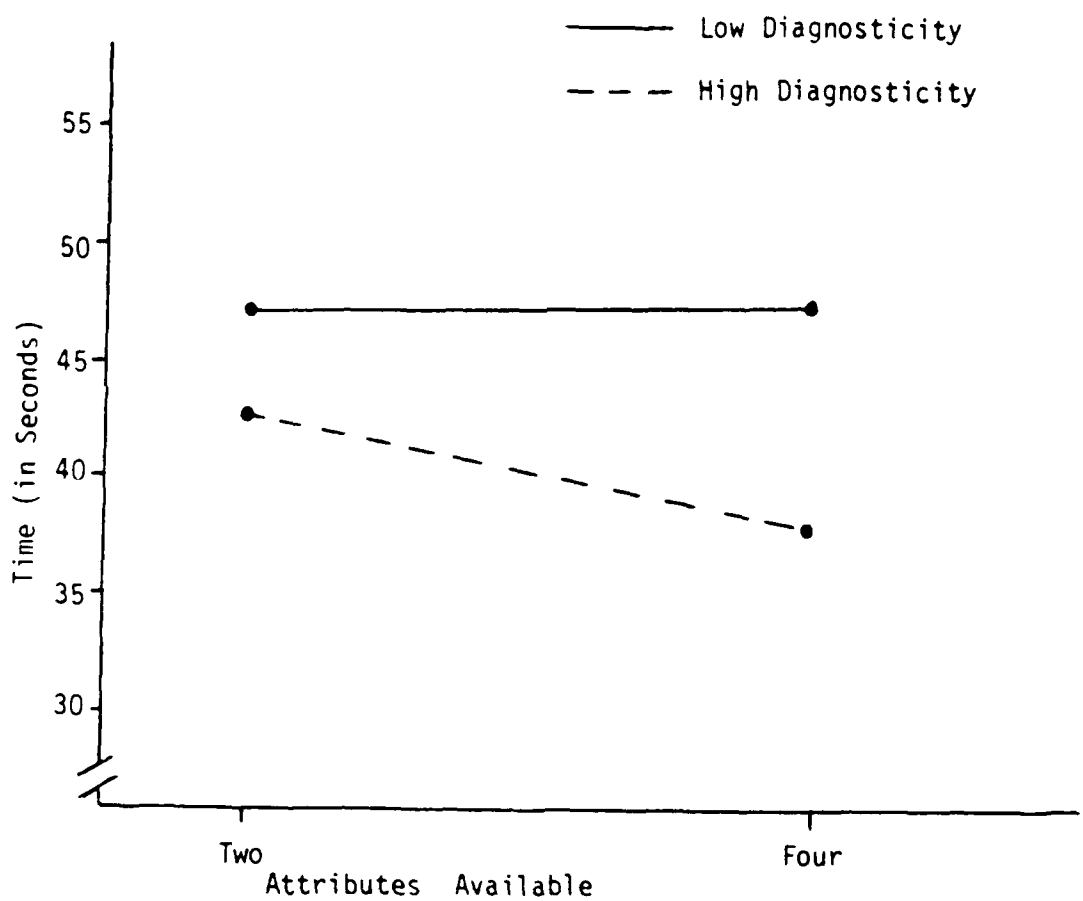


FIGURE 5. Mean Time to Perform the Task as a Function of Attributes Available and Diagnosticity

For the dependent variable of Number of Attributes Requested, experienced subjects requested significantly more attributes in Phase II (3.4) than they had in Phase I (3.0). In addition, both experienced and novice subjects requested more attributes for trials with Low Diagnosticity (3.7) than for trials with High Diagnosticity (3.1). Finally, for Phase II performance, a two-way interaction showed that the difficult low diagnosticity trials resulted in more attribute requests for both two and four attributes available (3.8 and 3.7, respectively), whereas the easier high diagnosticity trials resulted in more requests when there were more attributes available (3.3) than when there were only two available (2.9).

The remaining dependent variables included in the analyses concern the use of the expert system in Phase II (means were presented in Table 7). The first variable is the percentage of times that the expert was asked for "advice." This percentage was obtained by calculating the number of times out of four trials that the subject asked the expert. (These four trials consisted of four common and one rare animal.) Analyses showed that subjects asked the expert more often in the first session ($\bar{x}$=.50) than in the second session ($\bar{x}$=.36). In addition, subjects asked the expert more often for the low diagnosticity trials ($\bar{x}$=.64) than for the high diagnosticity trials ($\bar{x}$=.21). Finally, an interaction showed that the experienced subjects asked the expert LESS often for the high diagnosticity trials but MORE often for the harder low diagnosticity trials, as compared with the novice subjects (see Figure 6).

The second variable reflecting subjects' use of the expert was, given that the subject HAD asked the expert for advice, the percentage of trials in which the subject gave an answer that was the same as the one given by the expert; that is, how often the subject actually used the expert's advice. First,the expert's advice was used slightly over half of the time (.56) for session 1, but slightly under half of the time for session 2 (.44); this difference was significant at the .05 level. In addition, the subjects tended to rely on the expert answer more often for the difficult low diagnosticity trials ($\bar{x}$=.61) than for the easier high diagnosticity trials ($\bar{x}$=.39). There were no other variables which affected this measure, including the experience of the subjects.

The last variable to be discussed was not an objective measure, but rather, the subjective perception of the subjects. On the final questionnaire, subjects were asked to rate both themselves and the expert on a scale ranging from 1= extremely inaccurate to 20= perfect. These ratings were subjected to a 2 (experienced vs. novice) x 2 (self vs. expert) analysis of variance. Results showed that overall, subjects rated themselves as better on the task than the expert (mean ratings were 13.7 and 11.3, respectively), $F(1,22)= 6.3$, $p<.05$. However, the main effect cannot be interpreted independently because an interaction shows that this effect is due entirely to the perceptions of the experienced subjects. The experienced subjects strongly considered themselves to be better than the expert (mean self-rating was 14.3 and mean expert-rating was 9.7), whereas the novice subjects saw no difference between themselves and the expert (means were 13 and 13.1, respectively), $F(1,22)= 6.8$, $p=.01$. To demonstrate this effect in another

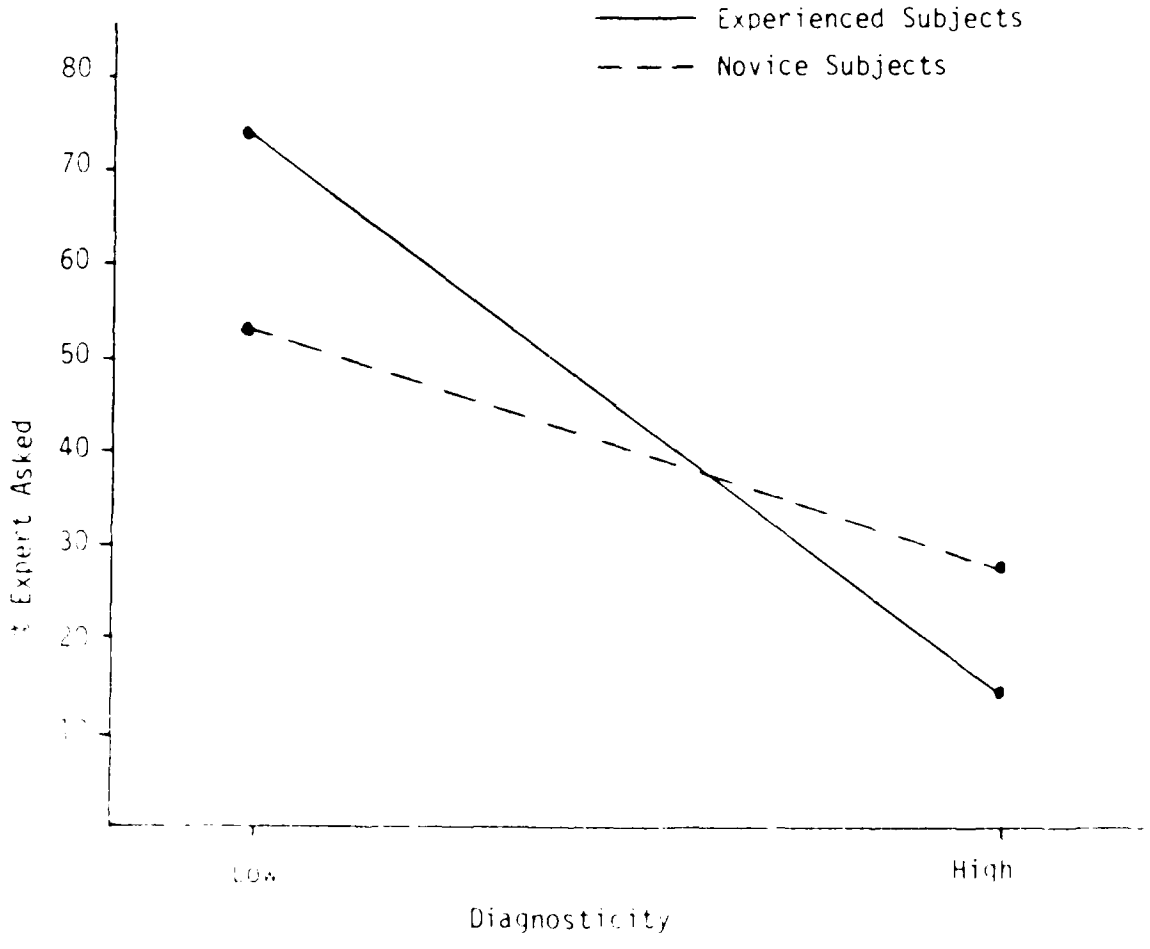


FIGURE 6. Mean % of Trials the Expert was Asked as a Function of Subject Experience and Diagnosticity

way, 10 out of the 12 experienced subjects rated themselves as being better than the expert, whereas only four of the novice subjects rated themselves as being better than the expert.

Strategy Analysis

Strategy analyses were conducted in a manner similar to that described for Phase I, using the same 12 subjects. Only four of the five strategies were assessed in this analysis; it was felt that the Hypothesis Testing strategy had been sufficiently ruled out and that the analysis should concentrate on the remaining four: Half-Split, Set Reduction, Favorite Attribute, and Random Request.

(1) Comparison of Half-Split, Set Reduction, and Random Request

To assess the likelihood of these three strategies, the frequencies were obtained for "best" split, "acceptable" split, and "non-diagnostic" first attribute requests. These data are listed in Table 8. As in the analysis for Phase I (see Table 2), these frequencies are totalled at the end of the table and can be compared with the expected frequencies based on the Half-Split, Set Reduction, and Random Request strategies. As in the Phase I analysis, the obtained values were closest to those predicted by the Set Reduction strategy. A Chi-Square analysis showed the frequencies to be significantly different from those predicted on the basis of random attribute request, Chi-Square(2)=72.8, $p < .001$.

To determine whether the request frequencies supported the Set Reduction or the Half-Split strategies, the frequencies were tallied for all conditions where a choice was made between the "best" split and "acceptable" split requests. The Half-Split strategy predicts that the "best" split request will always be chosen, whereas Set Reduction strategy predicts that there will be no difference between the two categories, and therefore the obtained frequencies should follow directly from the proportion of attributes in each of the two categories. The obtained values and the expected values based on the Set Reduction strategy were:

	"BEST" SPLIT	"ACCEPTABLE" SPLIT
OBTAINED	219	84
EXPECTED (Set Reduction)	177	126

Although the frequencies did not support the Half-Split strategy (i.e., all frequencies in the "best" split column), a Chi-Square test for goodness-of-fit showed that the data did not fit a Set Reduction strategy assuming EQUAL request of all diagnostic attributes, Chi-Square(1) =23.9, $p < .001$. The frequencies were biased in the direction of a Half-Split strategy. The most reasonable explanation of this finding is that the subjects were using Set Reduction, but were still somewhat more inclined to use attributes that evenly divide the alternative animals than attributes that provide a very uneven split.

Table 8. Percentage of First Attribute Requests--Categorized According to Type of Split (Phase II)

FIRST ATTRIBUTE	"BEST" SPLIT REQUESTS	"ACCEPTABLE" SPLIT REQUESTS	NON-DIAG REQUESTS
Small	Loc .32	Speed .02	
	Color .41	Noise .06	
	Alarm .19		
	TOTAL .92	TOTAL .08	---
Large	Speed .14		LOC .08
	Color .31		
	Noise .14		
	Alarm .33		
	TOTAL .92	--	TOTAL .08
Ground	Size .27	Noise .00	
	Speed .13		
	Color .33		
	Alarm .27		
	TOTAL 1.00	TOTAL .00	--
Tree	Color .50		Speed .03
	Noise .14		Size .03
	Alarm .30		
	TOTAL .94	--	TOTAL .06
Fast	Loc .15	Size .27	
	Color .33	Noise .09	
	Alarm .16		
	TOTAL .64	TOTAL .36	--
Slow	Size .50		Color .00
	Noise .33		Loc .00
	Alarm .17		
	TOTAL 1.00	--	TOTAL .00

Table 8. (Continued)

	"BEST" SPLIT REQUESTS		"ACCEPTABLE" SPLIT REQUESTS		NON-DIAG REQUESTS	
Brown	Size	.32				
	Loc	.36				
	Speed	.02				
	Noise	.10				
	Alarm	.20				
	TOTAL	1.00	--		--	
Grey	Size	.36			Noise	.07
	Loc	.38			Speed	.03
	Alarm	.16				
	TOTAL	.90	--		TOTAL	.10
No Noise	Color	.20	Loc	.24		
	Alarm	.20	Size	.32		
			Speed	.04		
	TOTAL	.40	TOTAL	.60	--	
Noise	Size	.36			Color	.14
	Loc	.21				
	Speed	.00				
	Alarm	.29				
	TOTAL	.86	--		TOTAL	.14
No Alarm	Size	.31				
	Loc	.17				
	Speed	.14				
	Color	.24				
	Noise	.14				
	TOTAL	1.00	--		--	
Alarm	Size	.44	Speed	.03		
	Loc	.25	Noise	.11		
	Color	.17				
	TOTAL	.86	TOTAL	.14	--	
Mean Percentage OBTAINED for all Combined (N=576):	(n=477)	.83	(n=84)	.14	(n=15)	.03

Table 8. (Concluded)

	"BEST" SPLIT REQUESTS	"ACCEPTABLE" SPLIT REQUESTS	NON-DIAG REQUESTS
Mean Percentage EXPECTED based on Split-Half Strategy:	1.00	.00	.00
Mean Percentage EXPECTED based on Set Reduction Strategy:	.78	.22	.00
Mean Percentage EXPECTED based on EQUAL Choice:	.67	.22	.11

An analysis was also conducted for the second attribute requests. As in the analysis for Phase I, these requests were divided into those where a choice was made between category #1 ("Best" Split) and category # 3 ("Non Diagnostic"), and those where a choice was made between category #1 ("Best" Split) and #2 ("Acceptable" Split). These data are presented in Table 9.

A Chi Square test was conducted for the "Best" Split vs. "Non-Diagnostic" requests where the obtained scores were compared with the frequencies expected on the basis of the Random Request strategy. The results showed that the data were significantly different from those expected on the basis of Random Request, Chi Square(2) =137, $p < .001$. It can be seen that relatively few choices were made for a non-diagnostic attribute.

A second Chi Square test was conducted for the "Best" Split vs "Acceptable" Split requests. In this case, the expected frequencies were provided by the Set Reduction strategy. The analysis showed that there was no significant difference between the values obtained and those expected on the basis of this model (Chi-Square < 3.0). These results provided direct support for the Set Reduction strategy.

Favorite Attributes

First and second attribute requests for Phase II were sorted to assess whether subjects had favorite attributes (as they did in Phase I), and also to determine whether this tendency changed from Phase I to Phase II.

Table 10 provides information similar to that given in Table 4; that is, the percentage of trials each subject asked for a favorite attribute, listed according to session. The mean percentages of favorite attributes are given at the bottom of the table, along with what would be expected if there were no preference in subjects' requests for the different attributes. In general, the tendency to rely on a favorite attribute increased from that found in Phase I, particularly for some of the subjects. A 2(Phase) x 2(Session) analysis of variance was performed on the percentage scores for the first attribute requests, and results showed that subjects did use one favorite attribute more often in their first request during Phase II ($\bar{x}$=.54) than in Phase I ($\bar{x}$=.40), $F(1,11)=10.8$, $p<.01$.

A similar analysis was performed for the second attribute requests. (This was performed separately because the percentages were based on fewer items than the first requests; subjects sometimes based their final decision on only two attributes.) The results of the analysis for second requests showed that although there was a slight trend towards a favorite more often in Phase II, this trend was not significant.

In summary, there was good support for the Set Reduction strategy, and there was also evidence not only that subjects were relying on favorite attributes within the confines of a Set Reduction strategy, but that this tendency increased in Phase II when the expert system was available for use.

Table 9. Frequency of Second Attribute Requests Categorized According to Type of Split (Phase II)

CATEGORY 1 VS. 3 CHOICE BETWEEN:	"BEST" SPLIT REQUESTS	"NON-DIAGNOSTIC" REQUESTS
3 Best, 1 Non-Diag.		
OBTAINED	90.0	8.0
EXPECTED	73.5	24.5
2 Best, 2 Non-Diag.		
OBTAINED	137.0	10.0
EXPECTED	73.5	73.5
1 Best, 3 Non-Diag.		
OBTAINED	22.0	4.0
EXPECTED	6.5	19.5
TOTAL OBTAINED	249.0	22.0
TOTAL EXPECTED (Random)	153.5	117.5
CATEGORY 1 VS. 2, CHOICE BETWEEN:		
3 Best, 1 Acceptable		
OBTAINED	47.0	13.0
EXPECTED	45.0	15.0
2 Best, 2 Acceptable		
OBTAINED	36.0	27.0
EXPECTED	31.5	31.5
1 Best, 3 Acceptable		
OBTAINED	13.0	23.0
EXPECTED	9.0	27.0
TOTAL OBTAINED	96.0	63.0
TOTAL EXPECTED (Set Reduction)	85.5	73.5

Table 10. Percentage of Favorite Attribute Requests for First and Second Requests (Phase II)

	FIRST REQUEST				SECOND REQUEST			
Subject	Session 2		Session 3		Session 2		Session 3	
1	Color	.58	Color	.42	Loc	.43	Color	.30
2	Loc	.37	Loc	.42	Color	.41	Color	.30
3	Size	.33	Size	.37	Size	.27	Size	.30
4	Noise	.75	Alarm	.83	Alarm	.62	Noise	.67
5	Size	.42	Size	.71	Noise	.27	Color	.29
6	Loc	.79	Loc	.62	Color	.62	Color	.39
7	Alarm	.54	Alarm	.54	Noise	.32	Size	.38
8	Size	.42	Size	.33	Alarm	.32	Alarm	.33
9	Speed	.21	Speed	.33	Speed	.32	Color	.25
10	Color	.42	Color	.87	Alarm	.30	Alarm	.42
11	Color	.75	Color	.87	Loc	.43	Loc	.48
12	Size	.46	Alarm	.54	Alarm	.29	Size	.27
Mean OBTAINED FREQUENCIES:		.50		.57		.38		.36
Mean EXPECTED FREQUENCIES (Random Choice):		.20		.20		.25		.25

IV. SUMMARY AND DISCUSSION

This section will first describe subject behavior in the inference task without the support of an Expert-Aiding system, and then describe how that behavior changes as a function of introducing an Expert-Aiding system. To accomplish this, the results will be summarized in two sections: The first will deal with overall performance of the task without and then with expert-aid, and the second will provide an analysis of cognitive strategies without and with expert-aid.

Performance

Manual Conditions

Subjects performing the task manually had no trouble learning the material or performing the inference task. Overall, subjects correctly solved an average of 73% of the trials. Subjects also found the task intrinsically interesting, and motivation was high. Thus, the development of the task as a research tool was considered quite successful, and will be utilized for extensions of the current research.

Several variables were manipulated to provide information bearing on the model given at the beginning of this report. It was felt that several factors would affect the subject's performance, WITHOUT consideration of whether there was a computer-aid available. The input variables which were assessed are given below on the right, with the corresponding manipulation provided on the left:

Manipulation	Model Variable
(1) Session	Experience
(2) Monetary Payoff	Seriousness of Consequences
(3) Attributes Available	Amount of Information Available
(4) Diagnosticity	Predictive Validity of Cue Set

(1) As expected, as the subjects' experience increased from session to session their accuracy improved significantly. Also, time to perform the task decreased. When subjects were asked to give a certainty rating for each guess, overall their certainty did not change over time. However, for the easy high diagnosticity tasks subjects became more certain of their answers, whereas for the difficult low diagnosticity tasks the certainty levels dropped. Also, with experience, subjects learned to ask for more information on the difficult trials.

(2) The second manipulation, Monetary Payoff, did not significantly affect any of the performance measures. This is interpreted as indicating that the operational definition of "seriousness of the consequences" was not sufficiently strong to have an effect on subjects. It is felt that the variable does have an impact in real-world inference tasks, but it is difficult to manipulate this variable in a laboratory setting.

(3 and 4) The last two variables, Attributes Available and Diagnosticity, are actually two factors which affect information transmitted to the subject. Under normal real-world circumstances, these variables would be correlated to some degree. However, in the present investigation the number of cues available and the diagnosticity of the cue set as a whole were varied orthogonally. The number of attributes available was either two or four, and the diagnosticity of the cue set was created to be either highly diagnostic (one possible answer) or of low diagnosticity (two or three answers possible).

As would be expected, the overall diagnosticity of the attribute set had a much greater effect on subjects' performance than did the number of attributes available (because if the subject acquired all relevant cues, the number of attributes available did not really affect the difficulty of the trial. In fact, Diagnosticity strongly impacted subject performance (mean performance level was .65 for the low diagnosticity trials and .87 for the high diagnosticity trials) while the number of attributes available did not affect performance. This indicates that, on the average, subjects did request information until the relevant information had been obtained. In addition to affecting performance, low diagnosticity caused subjects to take longer in performing the task, and subjects requested more information in these trials.

In Phase I, BOTH Diagnosticity and Attributes Available affected subject certainty ratings at the end of each trial. Subjects tended to be more confident on high diagnosticity trials than on low ones, but in addition, they were more confident when four attributes were available than when only two were available. This is obviously a subjective error in reasoning on the part of the subjects, because the number of attributes available did not directly affect the difficulty of the trial nor did it affect their performance.

Finally, a post-experimental questionnaire revealed that when subjects were asked to give estimates of their performance, they consistently underestimated their accuracy, regardless of experience (session).

To summarize results for Phase I, subjects became faster, more accurate, and more confident with experience, and the major factor affecting performance was the predictive validity (diagnosticity) of the attribute set. Subjects were more confident in their answers when they were given more information, although that variable did not affect their actual performance. Finally, when subjects were asked to estimate the accuracy of their performance, they consistently underestimated their performance.

Manual vs. Expert-Aiding Conditions

To assess the impact of introducing an expert-aid, subjects from the manual condition performed the task in two expert-aided sessions. In addition, new inexperienced (Novice) subjects were asked to perform the task under expert-aid conditions for two sessions. Thus, the experienced subjects' performance was compared with their own previous unaided performance as well as that of novice subjects who had not previously learned the task. The following variables were manipulated for this phase of the project:

Manipulation
(1) Manual vs. Expert-Aided
(2) Experienced vs. Novice
(3) Session
(4) Diagnosticity
(5) Attributes Available

(1 and 2) An important finding was that experienced subjects' accuracy declined from manual to expert-aided conditions. In addition, the novice subjects who performed only in Expert-Aided condition performed worse than the experienced subjects had performed during their first two sessions (see Figure 4). There are two plausible explanations for this detrimental effect. The first involves the fact that the Expert-Aid did not have information concerning the likelihood of each animal (some were common and others were rare). Since subjects did have access to this information, it is reasonable that their performance would be lower if they relied on the Expert-Aid (which evidence shows that they did). A second possible explanation for the decrease in performance is based on the fact that the actual task which subjects had to perform changed from Phase I to Phase II. In Phase I (manual conditions), subjects were asked to give a hypothesis and a certainty rating after the acquisition of EACH attribute. It is possible that this process forced the subjects to think more thoroughly about the task, and that this led to greater performance levels.

Data bearing on these alternative explanations come from a Phase x Diagnosticity interaction. Experienced subjects showed no difference between manual and Expert-Aided conditions for performance on the easy trials; however, their performance dropped considerably on the difficult trials. If the negative effect was due to task differences between the two phases, one would expect a drop in performance on ALL tasks. This finding, coupled with evidence that subjects asked and used the expert much more often for the difficult trials, lends credible support for the detrimental effect's being due to the use of the expert system. The important point in the above finding is not that subjects performed more poorly with the expert-aid, but that the experienced subjects DID rely on it even when they had been performing better on their own.

Another difference between the manual and expert-aided conditions was that subjects' certainty ratings after each guess were generally higher under the Expert-Aiding condition. However, the experienced subjects also had higher ratings than the novice subjects. The overall pattern of data indicates that it was experience and not the expert system that caused the subjects' ratings to increase. There was no evidence that subjects were more certain of their guesses simply because they could consult the advice of the expert system.

Subjects asked for more attribute information in the Expert-Aiding condition than in the manual condition. In addition, there was no difference between experienced and novice subjects. This suggests that something about the expert system caused subjects to obtain more attributes before making a guess. Again, there is a plausible alternative reason for this effect. In the manual condition, subjects were asked to make hypotheses and certainty

ratings after each attribute was provided. This may have caused impatience or frustration in performing the task, thereby reducing the amount of "cognitive effort" a subject wanted to put into any one trial. Thus, the overall effect might have resulted in a tendency to rely on fewer pieces of information in the manual condition. Research investigating these alternatives is currently being implemented.

The experienced and novice subjects were also compared for overall use of the expert system. First, use of the system by all subjects in general dropped from the first session (50%) to the second session (36%). There was no overall difference between frequency of use of the expert by experienced vs novice subjects. However, an interaction suggests that this finding is misleading. The experienced subjects used the expert more often for the difficult low diagnosticity trials and less often for the easy high diagnosticity trials, as compared with the novice subjects. The percentage of trials where the expert answer was used by subjects dropped from 56% for the first session to 44% for the second session. There were no differences between experienced and novice subjects in acceptance of the expert's answer.

Finally, a post-experimental questionnaire asked subjects to rate themselves and the expert. The experienced subjects rated themselves as being more accurate than the expert, whereas the novice subjects rated themselves as being equal to the expert.

(4 and 5) The characteristics of the trials were varied as before, by manipulating Diagnosticity of the total cue set and number of Attributes Available. The results followed the same pattern as in Phase I. For all subjects, the difficult low diagnosticity trials resulted in lower accuracy (Percent Correct), lower certainty ratings, and more attributes requested. As discussed previously, a Phase x Diagnosticity interaction showed that performance on the easy high diagnosticity trials remained relatively high when subjects went from manual to expert, whereas performance on the difficult low diagnosticity trials dropped when subjects used the expert system.

The number of attributes available did not affect performance in Phase I, but did slightly affect performance in Phase II (mean percent correct was .75 for two attributes available and .65 for four attributes available). This effect was opposite to what might be expected and was undoubtedly caused by subjects' not acquiring the information necessary on the four-attribute trials.

In summary, use of the expert system had a negative impact on subjects' accuracy, did not affect their certainty in making their guesses, and caused them to ask for more information than under manual conditions. Of particular importance, the experienced subjects showed better discrimination in using the expert system only for the most difficult trials and revealed in their ratings that they found it to be a less reliable system than their own abilities. On the other hand, novice subjects had never developed a good perception of the task and their own abilities, and so they used the expert system less discriminantly, and rated it equally as good as themselves.

Strategy Analysis for Manual Condition

To assess the strategies used by subjects in performing the task, five different possible strategies were first identified. In brief, these were: (1) Half-Split, the most rational and economical strategy, where subjects determine the set of possible animals based on the currently known set of attributes and then request another attribute which most optimally splits this set in half (this strategy is followed until one answer is determined or no attributes remain); (2) Set Reduction, a strategy where subjects think of SOME subset of possible answers based on the currently available set of attributes, and then request any attribute which will narrow down this subset (following this procedure means that the set of hypotheses may change substantially from one attribute acquisition to the next); (3) Hypothesis Testing, where subjects determine some subset of possible animals, with one obvious favorite, and then request an attribute that will confirm that favorite choice; (4) Favorite Attributes, where subjects do not attempt to reduce a set of alternatives but rather, have favorite attributes which they tend to request at first regardless of the initially given attribute; and (5) Random Request, an unlikely but possible strategy, where subjects simply randomly choose an attribute to request.

The strategies were compared by inspecting the patterns of attribute requests and also the hypothesized animals following each attribute acquisition. Several comparisons were made since there was no single analysis which could discriminate among all five strategies. In general, predictions were made for each strategy, and then data were sorted and analyzed using a Chi-Square goodness-of-fit test. All analyses effectively ruled out the Hypothesis Testing and Random Request strategies.

Two analyses of attribute requests were productive in discriminating between the Half-Split and Set Reduction strategies. The first attributes requested by subjects were analyzed in terms of predictions based on each of the two models, and the data clearly supported the Set Reduction strategy. That is, the subjects did not seem to mentally entertain all possible animals based on the initial attribute and then ask for an attribute which most evenly split the alternatives. Instead, they did think of some subset of alternatives and then ask for an attribute which would narrow the subset. This make intuitive sense, because the Set Reduction strategy requires much less cognitive energy and doesn't really cost the subject anything.

Analysis of the second attributes requested was interesting in that subjects seemed to follow the same Set Reduction strategy, but the data were "skewed" toward the Half-Split strategy. This means that, in general, the subjects tended to reduce the set; however, there was a definite indication that subjects were sometimes choosing an attribute which would best split the entire subset of possible animals. This is again intuitively plausible, because once subjects have two attributes, the entire subset of possible alternatives is small, and it is not difficult for them to mentally identify all alternatives at once.

A separate analysis of the attributes requested revealed that within the confines of the general strategy that subjects were using (Set Reduction) they did have some preference for particular attributes. That is, given that two attributes were both diagnostic, subjects would not randomly choose one but would request a preferred attribute. Also, some subjects showed this tendency more than others. This is presumably an effect of "availability," where some attributes are simply more available in memory than others. This would cause the subject to consider that attribute first, and if it was diagnostic, the subject would go ahead and request it.

In summary, the data show the general strategy to be that of Set Reduction, but as the task becomes less demanding, the subjects start requesting attributes that are more optimally diagnostic. The implications are that the greater the information load (number of possible attributes and number of possible causes), the greater the tendency to rely on the Set Reduction strategy (rather than the Half-Split). This would be especially important in a real-world task such as medical diagnosis, because the information load is great, and a tendency to mentally consider one small subset of causes (diseases) at first might induce a cognitive set that is never completely overcome. This is probably one area where computer systems may be most helpful (in making sure possible causes are not overlooked by the user). Finally, there seems to be an "availability" effect, where not all equally diagnostic attributes are requested equally often; rather, the subjects have favorites that are most accessible in memory, and if those attributes are diagnostic, they will be requested before any others.

Strategy Analysis for Expert-Aiding Condition

The same five strategies were assessed for subjects in the Expert-Aiding condition. The analyses were similar to those performed under manual conditions; attributes requested were compared with predictions based on the models, and Chi-square goodness-of-fit tests were performed. As before, the data were consistent with the Set Reduction strategy, with each subject showing some favoritism as far as requesting some attributes more than others. In fact, this tendency to request favorite attributes was more pronounced in the Expert-Aiding condition than under the manual condition. This is most likely because subjects were more inclined to "get information now" and "solve the problem later" with the expert system.

To summarize, subjects used what seems to be a sensible and relatively easy strategy, that of Set Reduction, in both the manual and Expert-Aiding conditions. Within the confines of that strategy, when there were several equally diagnostic attributes, subjects showed a tendency toward favoritism in requesting certain attributes more than others (different subjects favored different attributes). In addition, data indicate that as the set of possible alternatives and relevant attributes became small, subjects were more able to utilize a Half-Split strategy. However, the present study involved only two and four alternatives; in the real-world environment, this would occur only toward the end of the inference process.

V. CONCLUSIONS

Several useful pieces of information resulted from the present research. First, it was shown that the laboratory diagnostic inference task developed for the present effort is a successful tool for studying both manual performance of an inference task and the effects of introducing an expert-aiding system.

Second, preliminary comparisons of manual versus expert-aid performance of the diagnostic inference task have provided many insights into the impact of introducing an expert system. The most important of these is the benefits of training system operators to perform the task on their own, both to learn the task characteristics and also to become familiar with their own capabilities. This allows them to more effectively use and assess the expert system.

Finally, analysis of the strategies used by subjects in both the manual and Expert-Aiding condition revealed that subjects use an effective and cognitively non-demanding strategy in performing the inference task on their own, strategy that essentially does not change as a function of introducing an expert system.

There are obviously several details concerning the study which make these conclusions tentative until further research is conducted. Primarily, two issues need to be resolved. The first is whether the change in task requirements caused some of the differences found in the Expert-Aiding condition (such as the number of attributes requested). Second, the nature of the expert system needs to be modified such that it begins with at least as much information (probabilities, etc.) as the subject. This will make the system more accurate and also more closely mirror real-world expert systems. Research is currently being implemented to investigate these and other important issues.

REFERENCES

[1] Slovak, P., Fischoff, B., & Lichtenstein, S. (1977). Behavioral Decision Theory. Annual Review of Psychology, 28, 1-39.

[2] Price,I.E., Maisana, I.E., & Van Cott, H.P. (1982,June). The Allocation of Functions in Man-Machine Systems: A Perspective and Literature Review (NUREG CR-2623). Oak Ridge,TN: Oak Ridge National Laboratory.

[3] Yntema, D.B., & Torgerson, W.S. (1961). Man-Computer Cooperation in Decisions Requiring Common Sense. IRE Transactions on Human Factors in Electronics, 2, 20-26.

[4] Duda, R., Gaschnig, J., & Hart,O.E. (1981). Model Design in the Prospector Consultant System for Mineral Exploration. In B.L. Webber & N.J. Nilsson (Eds.), Readings in Artificial Intelligence. Palo Alto, CA: Tioga Publishing Co.

[5] Duda, R.O., & Gaschig, J.A. (1981). Knowledge-based Expert Systems Coming of Age. Byte, 6, 238-281.

[6] Shortliffe, E.H., Axline, S.G., Buchanan, B.G., Merigan, T.C., & Cohen, S.N. (1973). An Artificial Intelligence Program to Advise Physicians Regarding Antimicrobial Therapy. Computers in Biomedical Research, 6, 544-560.

[7] Shortliffe, E.H. Computer-Based Medical Consultant: MYCIN. New York: Elsevier/North Holland.

[8] Pope, H.E. (1981). Heuristic Methods for Imposing Structure on Ill-Structured Problems: The Structuring of Medical Diagnostics. In P. Szolovitz (Ed.), Artificial Intelligence in Medicine. Boulder,Co.: Westview Press.

[9] Buchanan, B.G., & Feigenbaum, E.A. DENDRAL and Meta-DENDRAL: Their Applications Dimension. Artificial Intelligence, 11, 5-24.

[10] Hillman, D.J. (1985). Artificial Intelligence. Human Factors, 27, 21-31.

[11] Brachman, R.J., & Smith, B.C. (1980). SIGART 70 (special issue on knowledge representation).

[12] Salvendy, G. (Ed.) (1984). Human-Computer Interaction. New York: Elsevier/North Holland.

[13] Michie, D. (1982). Introductory Readings in Expert Systems. New York: Gordon and Breach.

[14] Startsman, T.S., & Robinson, R.E. (1972). The Attitudes of Medical and Paramedical Personnel Towards Computers. Computers in Biomedical Research, 5B 218-227.

[15] Teach, R.L., & Shortliffe, E.H. (1981). An Analysis of Physician Attitudes Regarding Computer-Based Clinical Consultation Systems. Computers in Biomedical Research, 14, 542-558.

[16] Shortliffe, E.H. (1981). Medical Consultation Systems: Designing for Doctors. Designing for Human-Computer Communication. London: Academic Press.

[17] Price, H.E. (1985). The Allocation of Functions in Systems. Human Factors, 27B 33-45.

[18] Rouse, W.B. (in press). Models of Human Problem Solving: Detection, Diagnosis and Compensation for System Failures. Automatica.

[19] Rasmussen, J. (1979). On the Structure of Knowledge-A Morphology of Mental Models in a Man-Machine System Context (Report No. M-1983). Riso, Denmark: Riso National Laboratories.

[20] Shortliffe, E. H., Buchanan, B.G., & Feigenbaum, E.A. (1979). Knowledge Engineering for Medical Decision Making: A Review of Computer-based Clinical Decision Aids. Proceedings of the IEEE, 67,1207-1224.

[21] Fitter, M. J., & Cruikshank, P.J. (1893). Doctors Using Computers: A Case Study. In M.E. Sime & M.J. Coombs (Eds.), Designing for Human-Computer Communication. London: Academic Press.

[22] Hunt, R.M., & Rouse, W.B. (1981). Problem-Solving Skills of Maintenance Trainees in Diagnosing Faults in Simulated Power Plants. Human Factors, 23, 317-328.

APPENDIX A: POST EXPERIMENTAL QUESTIONNAIRES

A1 - Questionnaire for Manual Conditions:

1) What percentage of the trials do you think you answered correctly for each of the three sessions?

% for first day _____
% for second day _____
% for third day _____

2) For each of the eight animals, give the percentage of trials where the animal was the correct answer. That is, out of 100%, what percent accounted for each particular animal?

Rabbit _____
Groundhog _____
Deer _____
Bear _____
Squirrel _____
Owl _____
Hawk _____
Wolf _____

3) Try to describe the strategy you used during the trials.

A2 - Questionnaire for Expert-Aided Conditions:

1) On a scale from 1 to 20, rate yourself on how accurate you think YOUR OWN unaided guesses were (as you did each trial, how good were your ideas without considering the help you received). In rating yourself, 1 = extremely inaccurate and 20 = perfect.

2) On a scale from 1 to 20, rate the EXPERT on how accurate you think the answers it gave were, 1 = extremely inaccurate and 20 = perfect.

3) Was the expert better in some situations than others?

4) For each of the eight animals, give the percentage of trials where the animal was the correct answer. That is, out of 100%, what percent accounted for each particular animal?

Rabbit _____
Groundhog _____
Deer _____
Bear _____
Squirrel _____

Owl _____
Hawk _____
Wolf _____

5) Try to describe the strategy you used during the trials.

6) What percentage of the trials do you think you answered correctly for each of the two sessions?

% for first day _____
% for second day _____

APPENDIX B ANALYSIS OF VARIANCE TABLES PHASE I

Multivariate [illegible] for All Dependent Variables Combined:

[illegible]	[illegible] DF	Error DF	Mult F	Probability
Session	4	40	3.89	.01
Attribute AA	2	8	6.81	.01
Diagnosticity	4	8	43.74	.000
Session x AA	8	40	1.48	ns
Session x Diagnosticity	8	40	2.48	.03
AA x Diagnosticity	4	8	4.01	.04
Session x AA x Diag	8	40	1.97	ns

Analysis of Variance for PERCENT CORRECT:

Source	Source DF	Error DF	F	Probability
Session	2	22	7.58	.01
Attribute Avail (AA)	1	11	.07	ns
Diagnosticity	1	11	54.92	.000
Session x AA	2	22	1.82	ns
Session x Diag.	2	22	2.42	ns
AA x Diag.	1	11	.12	ns
Session x AA x Diag.	2	22	4.00	.03*

Analysis of Variance for TIME to perform the task:

Source	Source DF	Error DF	F	Probability
Session	2	22	19.02	.000
Attribute Avail (AA)	1	11	.18	ns
Diagnosticity	1	11	56.12	.000
Session x AA	2	22	1.75	ns
Session x Diag.	2	22	7.70	.01
AA x Diag.	1	11	2.87	ns
Session x AA x Diag.	2	22	2.20	ns

Analysis of Variance for Subjective CERTAINTY:

Source	Source DF	Error DF	F	Probability
Session	2	22	.19	ns
Attribute Avail. (AA)	1	11	23.60	.001
Diagnosticity	1	11	52.38	.000
Session x AA	2	22	1.20	ns
Session x Diag.	2	22	4.06	.03
AA x Diagnosticity	1	11	2.07	ns
Session x AA x Diag.	2	22	.38	ns

Analysis of Variance for Number of ATTRIBUTES REQUESTED:

Source	Source DF	Error DF	F	Probability
Session	2	22	.48	ns
Attribute Avail. (AA)	1	11	.73	ns
Diagnosticity	1	11	120.39	.000
Session x AA	2	22	.51	ns
Session x Diag.	2	22	4.30	.03
AA x Diagnosticity	1	11	.59	ns
Session x AA x Diag.	2	22	1.59	ns

Analysis of Variance for PERFORMANCE ESTIMATES:

Source	Source DF	Error DF	F	Probability
Session	2	22	9.22	.001
Estimated Vs. Actual	1	11	4.67	.05
Session x E/A	1	11	.55	ns

---

* Considered not significant because the Multivariate analysis was not significant for the three-way interaction

APPENDIX C: ANALYSIS OF VARIANCE TABLES FOR EXPERIENCED SUBJECTS, MANUAL VS. EXPERT-AID CONDITIONS

Multivariate Analysis for All Dependent Variables Combined:

Source	Source DF	Error DF	Mult. F	Prob.
Phase	4	8	28.92	.000
Session	4	8	4.11	.04
Attribute Av.(AA)	4	8	7.33	.01
Diagnosticity (Diag)	4	8	55.99	.000
Phase x Session	4	8	.30	ns
Phase x AA	4	8	1.53	ns
Phase x Diag.	4	8	6.59	.01
Session x AA	4	8	1.62	ns
Session x Diag.	4	8	1.33	ns
AA x Diag.	4	8	.95	ns
Phase x Session x AA	4	8	2.56	ns
Phase x Session x Diag.	4	8	.45	ns
Phase x AA x Diag.	4	8	.95	ns
Session x AA x Diag.	4	8	1.87	ns
Phase x Ses.x AA x Diag.	4	8	1.11	ns

Analysis of Variance for PERCENT CORRECT:

Source	Source DF	Error DF	F	Prob.
Phase	1	11	7.41	.02
Session	1	11	2.06	ns
Attribute Av.(AA)	1	11	2.39	ns
Diagnosticity (Diag)	1	11	155.46	.000
Phase x Session	1	11	.48	ns
Phase x AA	1	11	.48	ns
Phase x Diag.	1	11	5.98	.03
Session x AA	1	11	1.15	ns
Session x Diag.	1	11	1.74	ns
AA x Diag.	1	11	1.89	ns
Phase x Session x AA	1	11	1.64	ns
Phase x Session x Diag.	1	11	.34	ns
Phase x AA x Diag.	1	11	.46	ns
Session x AA x Diag.	1	11	4.07	ns
Phase x Ses. x AA x Diag.	1	11	.15	ns

Analysis of Variance for TIME to perform the task:

Source	Source DF	Error DF	F	Prob
Phase	1	11	46.91	.000*
Session	1	11	12.18	.01
Attribute Av.(AA)	1	11	7.92	.02
Diagnosticity (Diag.)	1	11	93.45	.000
Phase x Session	1	11	.09	ns

Phase x AA	1	11	.00	ns
Phase x Diag.	1	11	15.31	.01
Session x AA	1	11	2.22	ns
Session x Diag.	1	11	1.29	ns
AA x Diag.	1	11	1.56	ns
Phase x Session x AA	1	11	.01	ns
Phase x Session x Diag.	1	11	.09	ns
Phase x AA x Diag.	1	11	.59	ns
Session x AA x Diag.	1	11	6.01	.03**
Phase x Ses. x AA x Diag.	1	11	.72	ns

Analysis of Vraiance for Subjective CERTAINTY:

Source	Source DF	Error DF	F	Prob.
Phase	1	11	6.08	.03
Session	1	11	.10	ns
Attribute Available (AA)	1	11	18.02	.001
Diagnosticity (Diag.)	1	11	56.60	.000
Phase x Session	1	11	.92	ns
Phase x AA	1	11	3.30	ns
Phase x Diag.	1	11	.00	ns
Session x AA	1	11	1.53	ns
Session x Diag.	1	11	.69	ns
AA x Diag.	1	11	.54	ns
Phase x Session x AA	1	11	.83	ns
Phase x Session x Diag.	1	11	.79	ns
Phase x AA x Diag.	1	11	2.05	ns
Day x AA X Diag.	1	11	1.97	ns
Phase x Ses. x AA x Diag.	1	11	1.07	ns

Analysis of Variance for Number of ATTRIBUTES REQUESTED:

Source	Source DF	Error DF	F	Prob.
Phase	1	11	11.26	.01
Session	1	11	3.63	ns
Attribute Available (AA)	1	11	.43	ns
Diagnosticity (Diag.)	1	11	60.05	.000
Phase x Session	1	11	.01	ns
Phase x AA	1	11	.46	ns
Phase x Diag.	1	11	1.49	ns
Session x AA	1	11	4.05	ns
Session x Diag.	1	11	.05	ns
AA x Diag.	1	11	2.62	ns
Phase x Session x AA	1	11	1.27	ns
Phase x Session x Diag.	1	11	.45	ns

Phase x AA x Diag	1	11	1.48	ns
Session x AA x Diag.	1	11	.19	ns
Phase x Ses. x AA x Diag.	1	11	3.52	ns

---

* This variable is not considered because the subtasks required changed from Phase I to Phase II, affecting the time variable.

** Considered not significant because the multivariate analysis was not significant for this three-way interaction.

APPENDIX D: ANALYSIS OF VARIANCE TABLES FOR EXPERIENCED VS. NOVICE SUBJECTS, PHASE II

Multivariate Analysis for All Dependent Variables Combined:

Source	Source DF	Error DF	Mult. F	Prob.
Experience vs. Novice	6	17	2.60	.05
Session	6	17	5.72	.002
Attribute Av (AA)	6	17	3.60	.02
Diagnosticity (Diag.)	6	17	77.99	.000
E/N x Session	6	17	.85	ns
E/N x AA	6	17	.41	ns
E/N x Diag.	6	17	5.10	.01
Session x AA	6	17	2.57	ns
Session x Diag.	6	17	4.24	.01
AA x Diag.	6	17	3.68	.02
E/N x Session x AA	6	17	1.01	ns
E/N x Session x Diag.	6	17	1.37	ns
E/N x AA x Diag.	6	17	1.16	ns
Session x AA x Diag.	6	17	1.99	ns
E/N x Session x AA x Diag.	6	17	.97	ns

Analysis of Variance for PERCENT CORRECT:

Source	Source DF	Error DF	F	Prob.
Experience vs. Novice	1	22	4.19	.05
Session	1	22	.10	ns
Attribute Av(AA)	1	22	9.68	.01
Diagnosticity (Diag.)	1	22	125.22	.000
E/N x Session	1	22	.10	ns
E/N x AA	1	22	.06	ns
E/N x Diag.	1	22	.15	ns
Session x AA	1	22	.01	ns
Session x Diag.	1	22	4.02	.05
AA x Diag.	1	22	.91	ns
E/N x Session x AA	1	22	.36	ns
E/N x Session x Diag.	1	22	1.68	ns
E/N x AA x Diag.	1	22	2.17	ns
Session x AA x Diag.	1	22	8.96	.01*
E/N x Session x AA x Diag.	1	22	1.73	ns

Analysis of Variance for TIME to perform the task:

Source	Source DF	Error DF	F	Prob.
Experience vs. Novice	1	22	.61	ns
Session	1	22	29.43	.000
Attribute Avail.	1	22	4.43	.05

Diagnosticity (Diag.)	1	22	39.95	.000
E/N x Session	1	22	3.48	.07
E/N x AA	1	22	.60	ns
E/N x Diag.	1	22	2.47	ns
Session x AA	1	22	.26	ns
Session x Diag.	1	22	.26	ns
AA x Diag.	1	22	5.15	.03
E/N x Session x AA	1	22	.79	ns
E/N x Session x Diag.	1	22	.28	ns
E/N x AA x Diag.	1	22	.06	ns
Session x AA x Diag.	1	22	.38	ns
E/N x Session x AA x Diag.	1	22	.14	ns

Analysis of Variance for Subjective CERTAINTY:

Source	Source DF	Error DF	F	Prob.
Experience vs. Novice	1	22	10.23	.01
Session	1	22	.67	ns
Attribute Av(AA)	1	22	2.36	ns
Diagnosticity (Diag.)	1	22	108.28	.000
E/N x Session	1	22	.24	ns
E/N x AA	1	22	.34	ns
E/N x Diag.	1	22	.54	ns
Session x AA	1	22	.05	ns
Session x Diag.	1	22	11.71	.002
AA x Diag.	1	22	.06	ns
E/N x Session x AA	1	22	1.20	ns
E/N x Session x Diag.	1	22	3.97	.06
E/N x AA x Diag.	1	22	.01	ns
Session x AA x Diag.	1	22	.08	ns
E/N x Session x AA x Diag.	1	22	4.71	.04*

Analysis of Variance for Number of ATTRIBUTES REQUESTED:

Source	Source DF	Error DF	F	Prob.
Experience vs. Novice	1	22	.06	ns
Session	1	22	1.50	ns
Attribute Av(AA)	1	22	2.91	ns
Diagnosticity (Diag.)	1	22	30.47	.000
E/N x Session	1	22	.14	ns
E/N x AA	1	22	.38	ns
E/N x Diag.	1	22	2.81	ns
Session x AA	1	22	8.85	.01
Session x Diag.	1	22	.86	ns
AA x Diag.	1	22	10.60	.004
E/N x Session x AA	1	22	.01	ns
E/N x Session x Diag.	1	22	.07	ns
E/N x AA x Diag.	1	22	.18	ns
Session x AA x Diag.	1	22	1.56	ns
E/N x Session x AA x Diag.	1	22	.01	ns

Analysis of Variance for Number of times EXPERT was ASKED:

Source	Source DF	Error DF	F	Prob.
Experience vs. Novice	1	22	.24	ns
Session	1	22	9.47	.01
Attribute Av(AA)	1	22	3.83	.06
Diagnosticity (Diag.)	1	22	119.43	.000
E/N x Session	1	22	.00	ns
E/N x AA	1	22	1.47	ns
E/N x Diag.	1	22	18.99	.000
Session x AA	1	22	.30	ns
Session x Diag.	1	22	.99	ns
AA x Diag.	1	22	3.42	ns
E/N x Session x AA	1	22	1.44	ns
E/N x Session x Diag.	1	22	1.60	ns
E/N x AA x Diag.	1	22	5.43	.03*
Session x AA x Diag.	1	22	.18	ns
E/N x Session x AA x Diag.	1	22	.02	ns

Analysis of Variance for Number of times EXPERT was USED:

Source	Source DF	Error DF	F	Prob.
Experience vs. Novice	1	22	.43	ns
Session	1	22	5.24	.03
Attribute Available (AA)	1	22	.02	ns
Diagnosticity (Diag.)	1	22	14.69	.001
E/N x Session	1	22	.49	ns
E/N x AA	1	22	.66	ns
E/N x Diag.	1	22	.63	ns
Session x AA	1	22	.47	ns
Session x Diag.	1	22	.26	ns
AA x Diag.	1	22	.25	ns
E/N x Session x AA	1	22	.67	ns
E/N x Session x Diag.	1	22	.30	ns
E/N x AA x Diag.	1	22	2.08	ns
Session x AA x Diag.	1	22	.17	ns
E/N x Session x AA x Diag.	1	22	.73	ns

---

* Considered not significant because the multivariate analysis was not significant for this interaction.

END
DATE
FILMD
3-88
DTIC